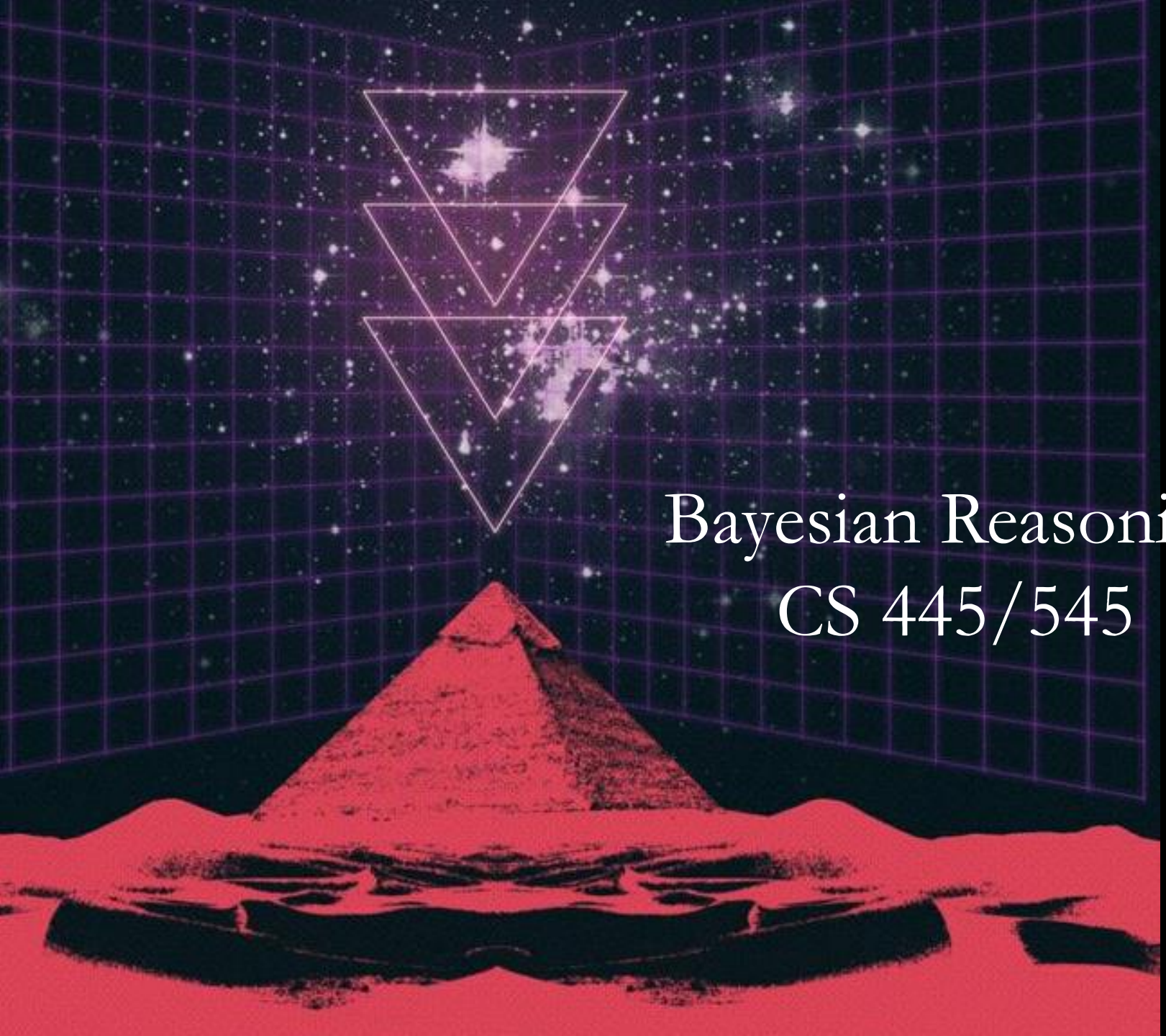
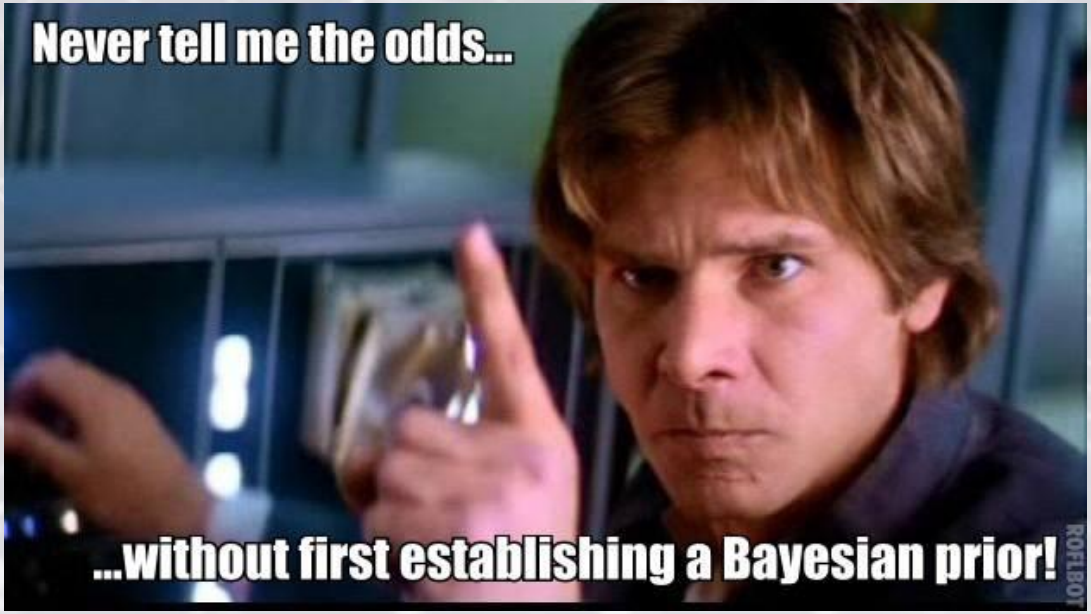


Bayesian Reasoning
CS 445/545

Bayesian and Frequentist Probability

- (2) General paradigms for statistics and statistical inference: *frequentist* vs. *Bayesian*.
- **Frequentists**: Parameters are fixed; there is a (Platonic) model; parameters remain constant.
- **Bayesians**: Data are fixed; data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.



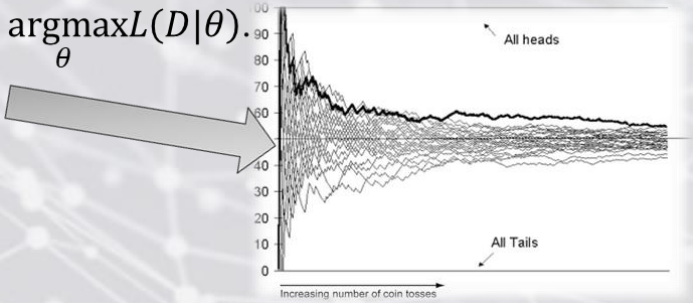
Bayesian and Frequentist Probability

- (2) General paradigms for statistics and statistical inference: *frequentist* vs. *Bayesian.*
- Frequentists: Parameters are fixed; there is a (Platonic) model; parameters remain constant.
- Bayesians: Data are fixed; data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.
- Frequentists commonly use the *MLE* (**maximum likelihood estimate**) as a cogent *point estimate*

of the model parameters of a probability distribution: $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} L(D|\theta)$.

- Using the *Law of Large Numbers (LLN),* $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$,

one can consequently show that: $\hat{\theta}_{MLE} \xrightarrow{P} \theta$.

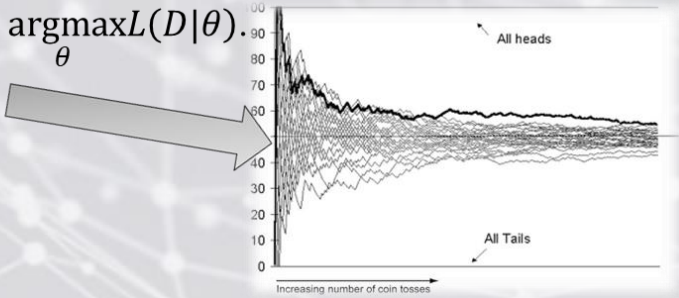


Bayesian and Frequentist Probability

- (2) General paradigms for statistics and statistical inference: *frequentist* vs. *Bayesian.*
- Frequentists: Parameters are fixed; there is a (Platonic) model; parameters remain constant.
- Bayesians: Data are fixed; data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.
- Frequentists commonly use the *MLE* (**maximum likelihood estimate**) as a cogent *point estimate*

  of the model parameters of a probability distribution: $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} L(D|\theta)$.
- Using the *Law of Large Numbers (LLN),* $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0,$

  one can consequently show that: $\hat{\theta}_{MLE} \xrightarrow{P} \theta.$



Potential issues with frequentist approach: philosophical reliance on long-term 'frequencies', *the problem of induction* (Hume) and the **black swan paradox**, as well as the presence of limited exact solutions for a small class of settings.

Bayesian and Frequentist Probability

• In the Bayesian framework, conversely, probability is regarded as a measure of uncertainty pertaining to the practitioner's knowledge about a particular phenomenon.

• The prior belief of the experimenter is not ignored but rather encoded in the process of calculating probability.

Bayesian and Frequentist Probability

• As the Bayesian gathers new information from experiments, this information is used, in conjunction with prior beliefs, to update the measure of certainty related to a specific outcome. These ideas are summarized elegantly in the familiar *Bayes' Theorem*:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

• Where *H* here connotes '*hypothesis*' and *D* connotes '*data*'; the leftmost probability is referred to as the *posterior* (of the hypothesis), and the numerator factors are called the *likelihood* (of the data) and the *prior* (on the hypothesis), respectively; the denominator expression is referred to as the *marginal likelihood*.

Bayesian and Frequentist Probability

As the Bayesian gathers new information from experiments, this information is used, in conjunction with prior beliefs, to update the measure of certainty related to a specific outcome. These ideas are summarized elegantly in the familiar *Bayes' Theorem*:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Where *H* here connotes '*hypothesis*' and *D* connotes '*data*'; the leftmost probability is referred to as the *posterior* (of the hypothesis), and the numerator factors are called the *likelihood* (of the data) and the *prior* (on the hypothesis), respectively; the denominator expression is referred to as the *marginal likelihood*.

Typically, the point estimate for a parameter used in Bayesian statistics is the *mode* of the *posterior distribution*, known as the **maximum a posterior** (MAP) estimate, which is given as:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)P(\theta)$$

Conditional Probability

- Probability of an event given the occurrence of some other event.

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

Deriving Bayes Rule

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

Bayes rule :

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Independence and Conditional Independence

- Recall that two random variables, X and Y, are **independent** if

$$P(X,Y) = P(X)P(Y)$$

- Two random variables, X and Y, are independent *given* C if

$$P(X,Y \mid C) = P(X \mid C)P(Y \mid C)$$

Inclusion-Exclusion Principle

- Probability of a *disjunction*:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- If a and b are *independent events*:

$$P(a \vee b) = P(a) + P(b)$$

Inclusion-Exclusion Principle

- Probability of a *disjunction*:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- The disjunction formula generalizes…

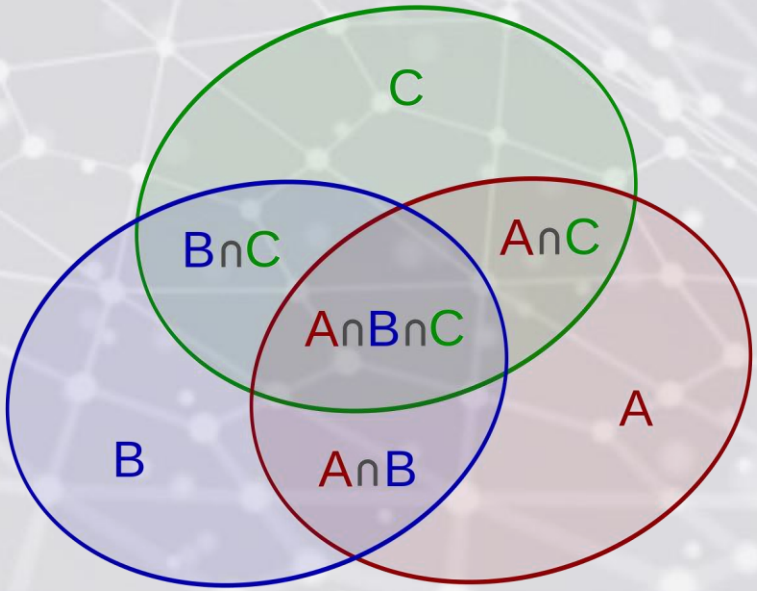
$$P(a \vee b \vee c) = ?$$

Inclusion-Exclusion Principle

- Probability of a *disjunction*:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

$$P(a \vee b \vee c) = P(a) + P(b) + P(c) - P(a \wedge b) - P(a \wedge c) - P(b \wedge c) + P(a \wedge b \wedge c)$$



General Application to Data Models

- In machine learning we have a space H of hypotheses:

  $h_1, h_2, \ldots, h_n$ (possibly infinite)

- We also have a set D of data

- We want to calculate $P(h \mid D)$

- Bayes rule gives us: $P(h \mid D) = \dfrac{P(D \mid h)P(h)}{P(D)}$

Terminology

- ***Prior probability* of *h*:**
  - $P(h)$: Probability that hypothesis h is true given our prior knowledge
  - If no prior knowledge, all $h \in H$ are equally probable
- ***Posterior probability* of *h*:**
  - $P(h \mid D)$: Probability that hypothesis h is true, given the data D.
- ***Likelihood* of *D*:**
  - $P(D \mid h)$: Probability that we will see data D, given hypothesis h is true.
- ***Marginal likelihood of D***
  - $P(D) = \sum_{h} P(D \mid h) P(h)$

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = .09/.2 = .45$$

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = .09/.2 = .45$$

Given that an email is spam, what is the probability it is in your junk folder?

Elementary Conditional Probability Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = .09/.2 = .45$$

Given that an email is spam, what is the probability it is in your junk folder?

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} = .09/.1 = .9$$

Testing the Validity of ESP

Your friend returns from a trip to New York City, reporting that he saw Madame Blavatsky, the famous clairvoyant, successfully predict the outcome of 100 coin tosses. Should we believe in ESP, the theory that some people have a magical ability to sense the future?

Testing the Validity of ESP

Your friend returns from a trip to New York City, reporting that he saw Madame Blavatsky, the famous clairvoyant, successfully predict the outcome of 100 coin tosses. Should we believe in ESP, the theory that some people have a magical ability to sense the future?

• Consider a set of "theories" *T*: {"ESP is real", "ESP is not real"} and a data set *D*: {"Madame B. is no better than chance at predicting the toss of an unbiased coin," "Madame Blavatsky can predict perfectly the outcome of 100 coin tosses."} – for simplicity we assume these are the only possible data.

We'll write: T={ESP, ~ESP}; D={normal, predict}, respectively.

Testing the Validity of ESP

Your friend returns from a trip to New York City, reporting that he saw Madame Blavatsky, the famous clairvoyant, successfully predict the outcome of 100 coin tosses. Should we believe in ESP, the theory that some people have a magical ability to sense the future?

• Consider a set of "theories" *T*: {"ESP is real", "ESP is not real"} and a data set *D*: {"Madame B. is no better than chance at predicting the toss of an unbiased coin," "Madame Blavatsky can predict perfectly the outcome of 100 coin tosses."} – for simplicity we assume these are the only possible data.

We'll write: T={ESP, ~ESP}; D={normal, predict}, respectively.

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = ?$$

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = \frac{P(predict \mid ESP)\,P(ESP)}{P(predict)}$$

P(predict|ESP) means "what's the chance that we get the data {Madame Blavatsky predicts perfectly} given the truth of the theory ESP?"

Let's say that if ESP is real, Madame Blavatsky almost certainly has it, and if she has it, she can do amazing predictions like these, so we set that at 0.9—i.e., only a 10% chance she'll screw up using her (real) magic powers.

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = \frac{P(predict \mid ESP)\,P(ESP)}{P(predict)}$$

P(predict|ESP) means "what's the chance that we get the data {Madame Blavatsky predicts perfectly} given the truth of the theory ESP."

Let's say that if ESP is real, Madame Blavatsky almost certainly has it, and if she has it, she can do amazing predictions like these, so we set that at 0.9—i.e., only a 10% chance she'll screw up using her (real) magic powers.

P(ESP) is the prior belief you have in ESP—the degree of belief you attribute to the possibility before hearing about the new data. Let's say you're a scientist; you attribute low value to these kinds of things, but (you're a scientist)—nothing is impossible, so we'll say 10−12. You're more confident that ESP is fake than you are about surviving your next airline flight.

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = \frac{P(predict \mid ESP)P(ESP)}{P(predict)}$$

Finally, P(predict): the probability this prediction event happens; recall that P(ESP|predict)+P(~ ESP|predict)=1.

$$P(predict) = P(predict \mid ESP)P(ESP) + P(predict \mid \sim ESP)P(\sim ESP)$$

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = \frac{P(predict \mid ESP)P(ESP)}{P(predict)}$$

Finally, P(predict): the probability this prediction event happens; recall that P(ESP|predict)+P(~ ESP|predict)=1.

$$P(predict) = P(predict \mid ESP)P(ESP) + P(predict \mid \sim ESP)P(\sim ESP)$$

$$P(ESP) = 10^{-12}, P(\sim ESP) = 1 - 10^{-12}$$

$$P(predict \mid ESP) = 0.9 \ (by\ assumption)$$

$$P(predict \mid \sim ESP) = 2^{-100}$$

Testing the Validity of ESP

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, "**conditional on predict, what degree of belief should I have in ESP**?"

$$P(ESP \mid predict) = \frac{P(predict \mid ESP)P(ESP)}{P(predict)}$$

Finally, P(predict): the probability this prediction event happens; recall that P(ESP|predict)+P(~ ESP|predict)=1.

$$P(predict) = P(predict \mid ESP)P(ESP) + P(predict \mid \sim ESP)P(\sim ESP)$$

$$P(ESP) = 10^{-12}, P(\sim ESP) = 1 - 10^{-12} \qquad P(predict \mid ESP) = 0.9 \ (by\ assumption)$$

$$P(predict \mid \sim ESP) = 2^{-100}$$

$$P(ESP \mid predict) = \frac{0.9 * 10^{-12}}{0.9 * 10^{-12} + 2^{-100}\left(1 - 10^{-12}\right)} \approx 1 - 10^{-18}$$

Thus, ESP is most certainly true, conditional on Madame B's ability to predict 100 consecutive coin flips.

Testing the Validity of ESP

$$P(ESP \mid predict) = \frac{0.9 * 10^{-12}}{0.9 * 10^{-12} + 2^{-100}\left(1 - 10^{-12}\right)} \approx 1 - 10^{-18}$$

Thus, ESP is most certainly true, conditional on Madame B's ability to predict 100 consecutive coin flips.

Or is it?

Let's redo the previous problem with an expanded theory set:

T= {"ESP is real, your friend is *not* delusional", "ESP is not real, your friend is *not* delusional', "ESP is real, your friend *is* delusional", "ESP is not real, your friend *is* delusional"}.

We'll abbreviate as before: {ESP&~D, ~ESP&~D, ESP&D, ~ESP&D}.

Testing the Validity of ESP

T= {"ESP is real, your friend is *not* delusional", "ESP is not real, your friend is *not* delusional', "ESP is real, your friend *is* delusional", "ESP is not real, your friend *is* delusional"}.

We'll abbreviate as before: {ESP & ~D, ~ESP & ~D, ESP & D, ~ESP & D}.

It's probably safe to assume that these events are independent, so that:

$$P(ESP \,\&\, D) = P(ESP)P(D)$$

Testing the Validity of ESP

Consider: $$P(ESP \,\&\, D \mid predict) = \frac{P(predict \mid ESP \,\&\, D)P(ESP)P(D)}{P(predict)}$$

If we compute this quantity directly using Bayes' rule, the calculation is tedious, as the denominator requires four individual terms.

Alternatively, we can compare the **odds-ratio** of ESP & ~D vs. ~ESP & D.

Testing the Validity of ESP

Consider: $$P(ESP \,\&\, D \mid predict) = \frac{P(predict \mid ESP \,\&\, D)P(ESP)P(D)}{P(predict)}$$

If we compute this quantity directly using Bayes' rule, the calculation is tedious, as the denominator requires four individual terms.

Alternatively, we can compare the **odds-ratio** of ESP & ~D vs. ~ESP & D. In other words: How much more likely is it that ESP is false, and your friend is delusional, rather than ESP is true and your friend is not delusional?

Testing the Validity of ESP

Consider: $$P(ESP \,\&\, D \mid predict) = \frac{P(predict \mid ESP \,\&\, D)P(ESP)P(D)}{P(predict)}$$

If we compute this quantity directly using Bayes' rule, the calculation is tedious, as the denominator requires four individual terms.

Alternatively, we can compare the **odds-ratio** of ESP & ~D vs. ~ESP & D. In other words: How much more likely is it that ESP is false, and your friend is delusional, rather than ESP is true and your friend is not delusional?

$$\frac{P(\sim ESP \,\&\, D \mid predict)}{P(ESP \,\&\, \sim D \mid predict)} = \frac{P(predict \mid \sim ESP \,\&\, D)P(\sim ESP)P(D)}{P(predict \mid ESP \,\&\, \sim D)P(ESP)P(\sim D)}$$

Testing the Validity of ESP

Alternatively, we can compare the **odds-ratio** of ESP & ~D vs. ~ESP & D. In other words: How much more likely is it that ESP is false, and your friend is delusional, rather than ESP is true and your friend is not delusional?

$$\frac{P(\sim ESP \,\&\, D \mid predict)}{P(ESP \,\&\sim D \mid predict)} = \frac{P(predict \mid \sim ESP \,\&\, D)P(\sim ESP)P(D)}{P(predict \mid ESP \,\&\sim D)P(ESP)P(\sim D)}$$

The only quantity we still need to specify is P(predict|~ESP & D). Let's say this value is 0.9, meaning the probability of perceived perfect prediction given that ESP is not real and your friend is delusional is quite high (naturally).

Testing the Validity of ESP

Alternatively, we can compare the **odds-ratio** of ESP & ~D vs. ~ESP & D. In other words: How much more likely is it that ESP is false, and your friend is delusional, rather than ESP is true and your friend is not delusional?

$$\frac{P(\sim ESP \,\&\, D \mid predict)}{P(ESP \,\&\sim D \mid predict)} = \frac{P(predict \mid \sim ESP \,\&\, D)P(\sim ESP)P(D)}{P(predict \mid ESP \,\&\sim D)P(ESP)P(\sim D)}$$

The only quantity we still need to specify is P(predict | ~ESP & D). Let's say this value is 0.9, meaning the probability of perceived perfect prediction given that ESP is not real and your friend is delusional is quite high (naturally).

$$\frac{P(\sim ESP \,\&\, D \mid predict)}{P(ESP \,\&\sim D \mid predict)} = \frac{P(predict \mid \sim ESP \,\&\, D)P(\sim ESP)P(D)}{P(predict \mid ESP \,\&\sim D)P(ESP)P(\sim D)} = \frac{0.9*(1-10^{-12})*10^{-6}}{0.9*10^{-12}*(1-10^{-6})} \approx 10^{6}$$

In other words: **it's a million times more likely that your friend is delusional, than it is that ESP is real**!

Testing the Validity of ESP

$$\frac{P(\sim ESP \,\&\, D \mid predict)}{P(ESP \,\&\sim D \mid predict)} = \frac{P(predict \mid \sim ESP \,\&\, D)P(\sim ESP)P(D)}{P(predict \mid ESP \,\&\sim D)P(ESP)P(\sim D)} = \frac{0.9*(1-10^{-12})*10^{-6}}{0.9*10^{-12}*(1-10^{-6})} \approx 10^{6}$$

In other words: **it's a million times more likely that your friend is delusional, than it is that ESP is real**!

D. Hume, "of Miracles", from an **Enquiry*… (1748):

> The plain consequence is (and it is a general maxim worthy of our attention), "That no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavours to establish: And even in that case, there is a mutual destruction of arguments, and the superior only gives us an assurance suitable to that degree of force, which remains, after deducting the inferior."
>
> When anyone tells me, that he saw a dead man restored to life, I immediately consider with myself, whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority, which I discover, I pronounce my decision, and always reject the greater miracle. If the falsehood of his testimony would be more miraculous, than the event which he relates; then, and not till then, can he pretend to command my belief or opinion.

*Recommended reading: "An Enquiry Concerning Human Understanding," Hume, 1748.

On the Matter of Priors

• A prior probability of an uncertain quantity is the probability distribution that would express one's **beliefs about this quantity before some evidence is taken into account**.

(*) Priors can be created using a myriad of methods. A prior can be determined from **past information**, such as previous experiments. A prior can also be *elicited* from the **purely subjective assessment of an experienced expert**.

On the Matter of Priors

• A prior probability of an uncertain quantity is the probability distribution that would express one's **beliefs about this quantity before some evidence is taken into account**.

(*) Priors can be created using a myriad of methods. A prior can be determined from **past information**, such as previous experiments. A prior can also be *elicited* from the **purely subjective assessment of an experienced expert**.

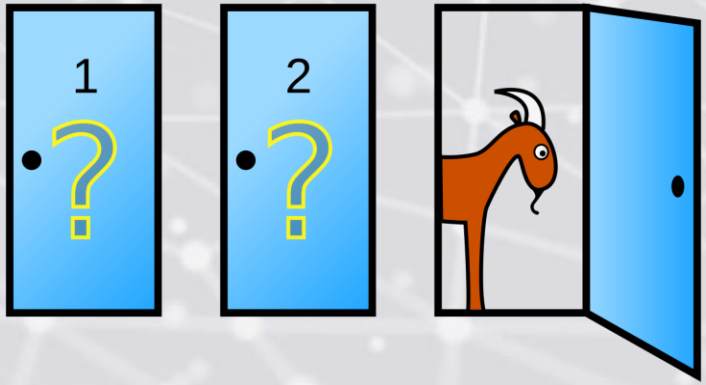
(*) **The principle of insufficient reason (PIR, Jackob Bernoulli, Laplace)** (also called: the principle of indifference, Keynes) states that if we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way.

Digression: The Monty Hall Problem

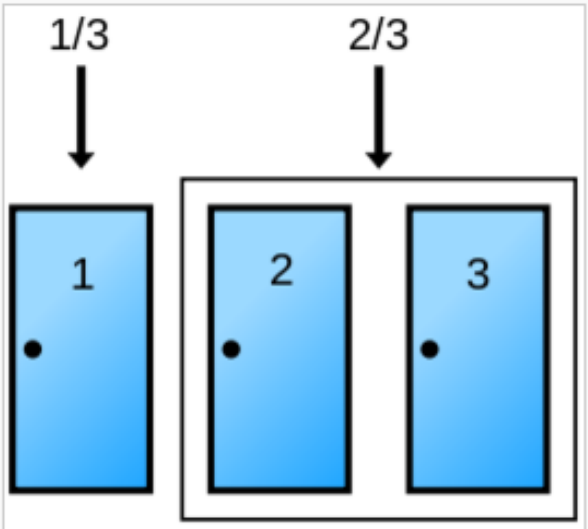
- Suppose you're on a game show, and you're given the choice of three doors:

Behind one door is a car; behind the others, goats.

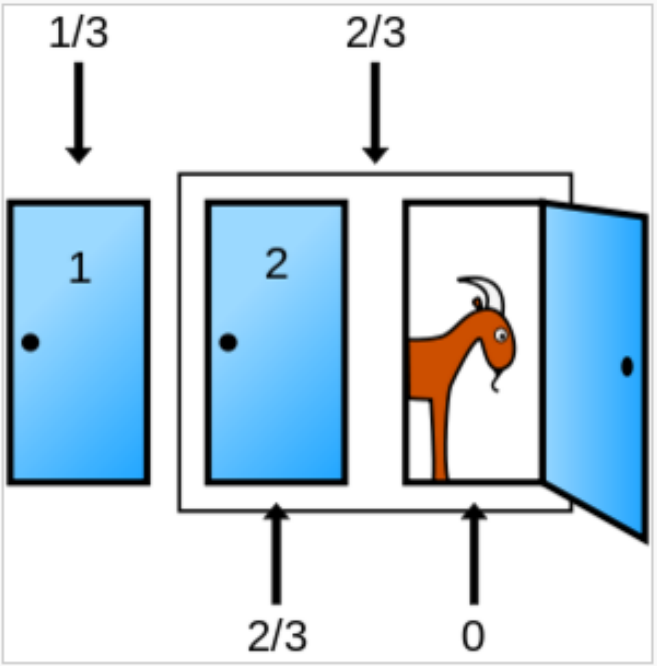
You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



Digression: The Monty Hall Problem

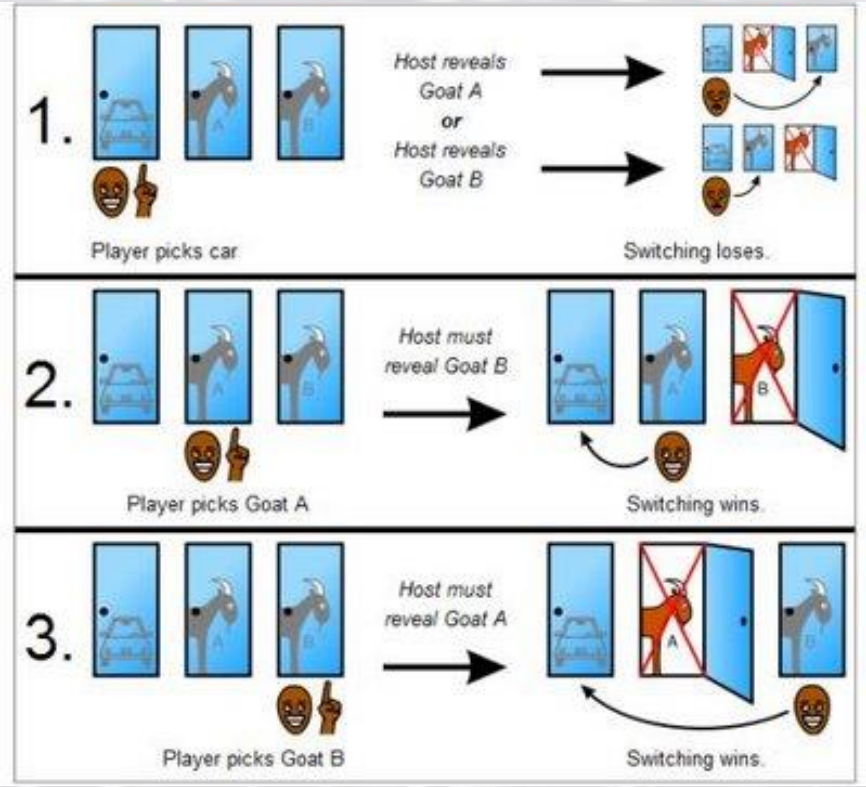


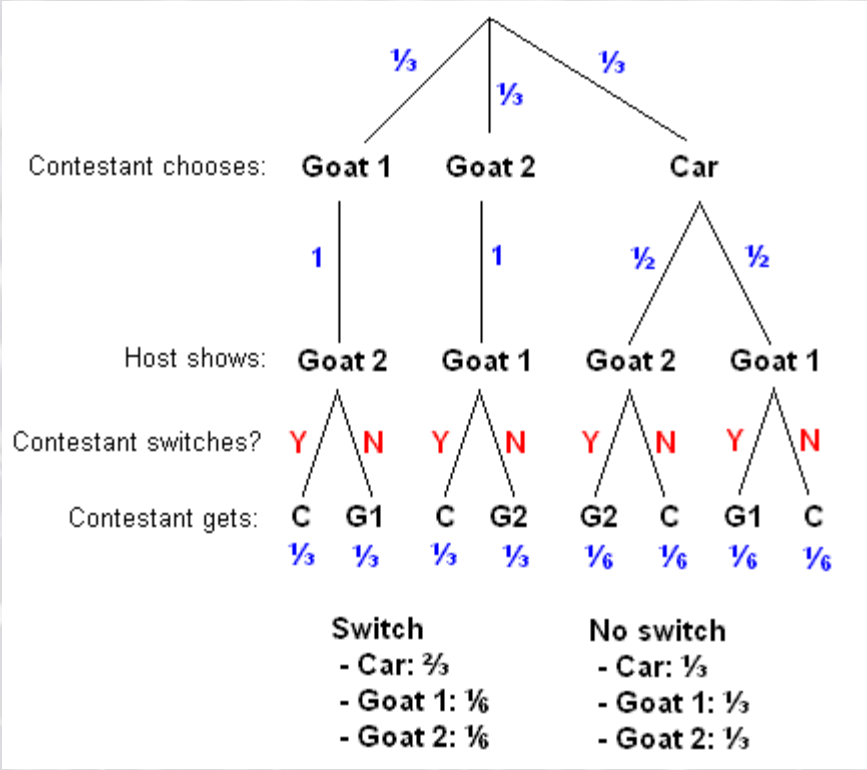
Car has a 1/3 chance of being behind the player's pick and a 2/3 chance of being behind one of the other two doors.



The host opens a door, the odds for the two sets don't change but the odds move to 0 for the open door and 2/3 closed door.

Digression: The Monty Hall Problem





Bayesian probability formulation

Hypothesis space *H*:

h_1 = Car is behind door A

h_2 = Car is behind door B

h_3 = Car is behind door C

Data *D*: After you picked door A, Monty opened B to show a goat

What is $P(h_1 \mid D)$?

What is $P(h_2 \mid D)$?

What is $P(h_3 \mid D)$?

Prior probability:

$P(h_1) = 1/3$ $P(h_2) = 1/3$ $P(h_3) = 1/3$

Likelihood:

$P(D \mid h_1) = 1/2$

$P(D \mid h_2) = 0$

$P(D \mid h_3) = 1$

Marginal likelihood:

$$P(D) = p(D|h_1)p(h_1) + p(D|h_2)p(h_2) + p(D|h_3)p(h_3) = 1/6 + 0 + 1/3 = 1/2$$

By Bayes rule:

$$P(h_1 \mid D) = \frac{P(D \mid h_1)P(h_1)}{P(D)} = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)(2) = \frac{1}{3}$$

$$P(h_2 \mid D) = \frac{P(D \mid h_2)P(h_2)}{P(D)} = (0)\left(\frac{1}{3}\right)(2) = 0$$

$$P(h_3 \mid D) = \frac{P(D \mid h_3)P(h_3)}{P(D)} = (1)\left(\frac{1}{3}\right)(2) = \frac{2}{3}$$

So you should switch!

MAP ("maximum a posteriori") Learning

Bayes rule: $P(h \mid D) = \dfrac{P(D \mid h)P(h)}{P(D)}$

Goal of learning: Find maximum a posteriori hypothesis h_{MAP}:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h \mid D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D \mid h)P(h)$$

because $P(D)$ is a constant independent of h.

Note: If every $h \in H$ is equally probable, then

$$h_{\text{MAP}} = \underset{h \in H}{\operatorname{argmax}} \; P(D \mid h)$$

h_{MAP} is called the “maximum likelihood hypothesis”.

A Medical Example

Toby takes a test for leukemia. The test has two outcomes: positive and negative. It is known that if the patient has leukemia, the test is positive 98% of the time. If the patient does not have leukemia, the test is positive 3% of the time. It is also known that 0.008 of the population has leukemia.

Toby's test is positive.

Which is more likely: Toby has leukemia or Toby does not have leukemia?

- **Hypothesis space:**

  h_1 = T. has leukemia

  h_2 = T. does not have leukemia

- **Prior:** 0.008 of the population has leukemia. Thus

  $P(h_1) = 0.008$

  $P(h_2) = 0.992$

- **Likelihood:**

  $P(+ \mid h_1) = 0.98$, $P(- \mid h_1) = 0.02$

  $P(+ \mid h_2) = 0.03$, $P(- \mid h_2) = 0.97$

- **Posterior knowledge:**

  Blood test is + for this patient.

- In summary

$P(h_1) = 0.008,\ P(h_2) = 0.992$

$P(+ \mid h_1) = 0.98,\ P(- \mid h_1) = 0.02$

$P(+ \mid h_2) = 0.03,\ P(- \mid h_2) = 0.97$

- Thus:

$$h_{MAP} = \underset{h \in H}{argmax}\ P(D \mid h)P(h)$$

$$P(+ \mid leukemia)P(leukemia) = (0.98)(0.008) = 0.0078$$

$$P(+ \mid \text{Ø}leukemia)P(\text{Ø}leukemia) = (0.03)(0.992) = 0.0298$$

$$h_{MAP} = \text{Ø}leukemia$$

- What is P(leukemia|+)?

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

So, $$P(leukemia \mid +) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\varnothing leukemia \mid +) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

These are called the “posterior” probabilities.

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+1, -1\}$.

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$

We want to find the most probable class value, h_{MAP}, given the data $\mathbf{x}$:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\text{argmax}} P(class \mid D)$$

$$= \underset{class \in \{+1,-1\}}{\text{argmax}} P(class \mid x_1, x_2, ..., x_n)$$

By Bayes Theorem:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\text{argmax}} \frac{P(x_1, x_2, ..., x_n \mid class)P(class)}{P(x_1, x_2, ..., x_n)}$$

$$= \underset{class \in \{+1,-1\}}{\text{argmax}} P(x_1, x_2, ..., x_n \mid class)P(class)$$

$P(class)$ can be estimated from the training data. How?

However, in general, not practical to use training data to estimate $P(x_1, x_2, ..., x_n \mid class)$. Why not?

- Naive Bayes classifier: Assume

$$P(x_1, x_2, ..., x_n \mid class) = P(x_1 \mid class)P(x_2 \mid class)\cdots P(x_n \mid class)$$

(*) In other words, with the **naïve Bayes classifier**, we assume the features are conditionally independent, given the class.

Is this a good assumption?

- Naive Bayes classifier: Assume

$$P(x_1, x_2, ..., x_n \mid class) = P(x_1 \mid class)P(x_2 \mid class)\cdots P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = (x_1, x_2, ..., x_n)$:

Naive Bayes classifier:

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class)\prod_i P(x_i \mid class)$$

To train: Estimate the values of these various probabilities over the training set.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Test data:

D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Use training data to compute a probabilistic *model*:

$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$

$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$

$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$

$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$

$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$

$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$

$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$

$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$

$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$

$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)$$

Estimating probabilities / Smoothing

- **Recap:** In previous example, we had a training set and a new example,
  (Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong)

- We asked: What classification is given by a naive Bayes classifier?

- Let n_c be the number of training instances with class c.

- Let $n_c^{x_i=a_k}$ be the number of training instances with attribute value $x_i=a_k$ and class c.

Then: $$P(x_i = a_i \mid c) = \frac{n_c^{x_i=a_k}}{n_c}$$

- **Problem with this method:** If n_c is very small, gives a poor estimate.

- E.g., $P(Outlook = Overcast \mid no) = 0$.

- Now suppose we want to classify a new instance:

(Outlook=overcast, Temperature=cool, Humidity=high, Wind=strong)

Then: $$P(\text{no}) \tilde{O}_i P(x_i \mid \text{no}) = 0$$

This incorrectly gives us zero probability due to small sample.

One solution: *Laplace smoothing** (also called "add-one" smoothing)

For each class *c* and attribute x_i with value a_k, add one "virtual" instance.

That is, for each class c, recalculate:

$$P(x_i = a_i \mid c) = \frac{n_c^{x_i=a_k} + 1}{n_c + K}$$

where *K* is the number of possible values of attribute *a*.

***NB**: Laplace smoothing can be derived using Bayesian statistics by using a *multinomial likelihood* and an *uninformative Dirichlet prior*.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Laplace smoothing: Add the following virtual instances for *Outlook*:

Outlook=Sunny: **Yes** *Outlook=Overcast*: **Yes** *Outlook=Rain*: **Yes**

Outlook=Sunny: **No** *Outlook=Overcast*: **No** *Outlook=Rain*: **No**

$$P(Outlook = overcast \mid \mathbf{No}) = \frac{0}{5} \rightarrow \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{0+1}{5+3} = \frac{1}{8}$$

$$P(Outlook = overcast \mid \mathbf{Yes}) = \frac{4}{9} \rightarrow \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{4+1}{9+3} = \frac{5}{12}$$

$$P(Outlook = Sunny \mid Yes) = 2/9 \rightarrow 3/12 \quad P(Outlook = Sunny \mid No) = 3/5 \rightarrow 4/8$$

$$P(Outlook = Overcast \mid Yes) = 4/9 \rightarrow 5/12 \quad P(Outlook = Overcast \mid No) = 0/5 \rightarrow 1/8$$

$$P(Outlook = Rain \mid Yes) = 3/9 \rightarrow 4/12 \quad P(Outlook = Rain \mid No) = 2/5 \rightarrow 3/8$$

$$P(Humidity = High \mid Yes) = 3/9 \rightarrow 4/11 \quad P(Humidity = High \mid No) = 4/5 \rightarrow 5/7$$

$$P(Humidity = Normal \mid Yes) = 6/9 \rightarrow 7/11 \quad P(Humidity = Normal \mid No) = 1/5 \rightarrow 2/7$$

Etc.

Naive Bayes on continuous-valued attributes

- How to deal with continuous-valued attributes?

Two possible solutions:

- Discretize
- Assume particular probability distribution of classes over values (estimate parameters from training data)

Discretization: Equal-Width Binning

For each attribute x_i, create k equal-width bins in interval from $min(x_i)$ to $max(x_i)$.

The discrete "attribute values" are now the bins.

Questions: What should k be? What if some bins have very few instances?

Problem with balance between *discretization bias* and *variance*.

The more bins, the lower the bias, but the higher the variance, due to small sample size.

Discretization: Equal-Frequency Binning

For each attribute x_i , create k bins so that each bin contains an equal number of values.

Also has problems: What should k be? Hides outliers. Can group together instances that are far apart.

Gaussian Naïve Bayes

Assume that within each class, values of each numeric feature are normally distributed:

$$p(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu_{i,c}$ is the mean of feature *i* given the class c, and $\sigma_{i,c}$ is the standard deviation of feature *i* given the class c

We estimate $\mu_{i,c}$ and $\sigma_{i,c}$ from training data.

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

$P(\mathbf{POS}) = 0.5$

$P(\mathbf{NEG}) = 0.5$

$$\mu_{1,\mathbf{POS}} = \frac{(3.0 + 4.1 + 7.2)}{3} = 4.8$$

$$\sigma_{1,\mathbf{POS}} = \sqrt{\frac{(3.0 - 4.8)^2 + (4.1 - 4.8)^2 + (7.2 - 4.8)^2}{3}} = 1.8$$

$$\mu_{1,\mathbf{NEG}} = \frac{(2.0 + 4.1 + 8.1)}{3} = 4.7$$

$$\sigma_{1,\mathbf{NEG}} = \sqrt{\frac{(2.0 - 4.7)^2 + (4.1 - 4.7)^2 + (8.1 - 4.7)^2}{3}} = 2.5$$

$$\mu_{2,\mathbf{POS}} = \frac{(5.1 + 6.3 + 9.8)}{3} = 7.1$$

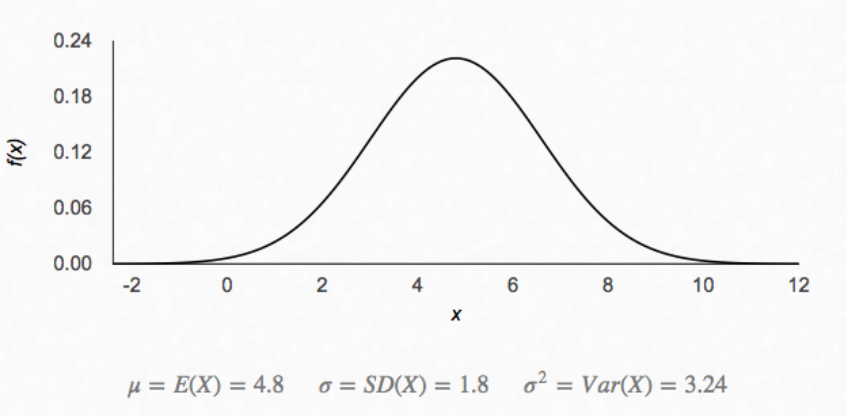
$$\sigma_{2,\mathbf{POS}} = \sqrt{\frac{(5.1 - 7.1)^2 + (6.3 - 7.1)^2 + (9.8 - 7.1)^2}{3}} = 2.0$$

$$\mu_{2,\mathbf{NEG}} = \frac{(1.1 + 2.0 + 9.4)}{3} = 4.2$$

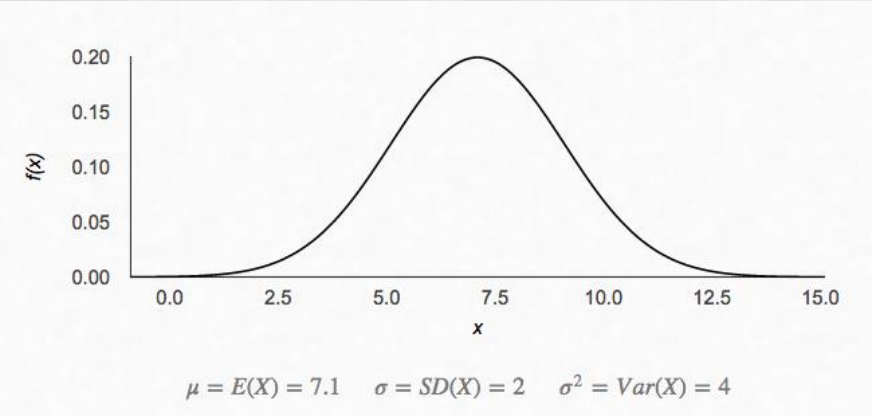
$$\sigma_{2,\mathbf{NEG}} = \sqrt{\frac{(1.1 - 4.2)^2 + (2.0 - 4.2)^2 + (9.4 - 4.2)^2}{3}} = 3.7$$

http://homepage.stat.uiowa.edu/~mbognar/applets/normal.html

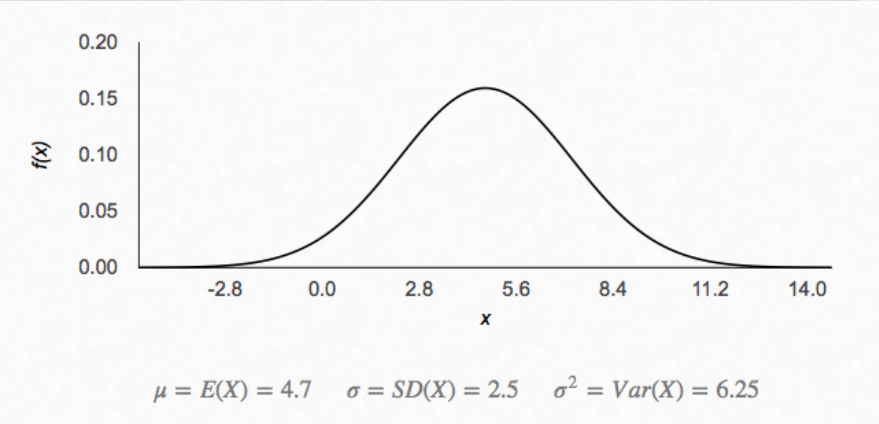
$N_{1,\mathbf{POS}} = N(x; 4.8, 1.8)$



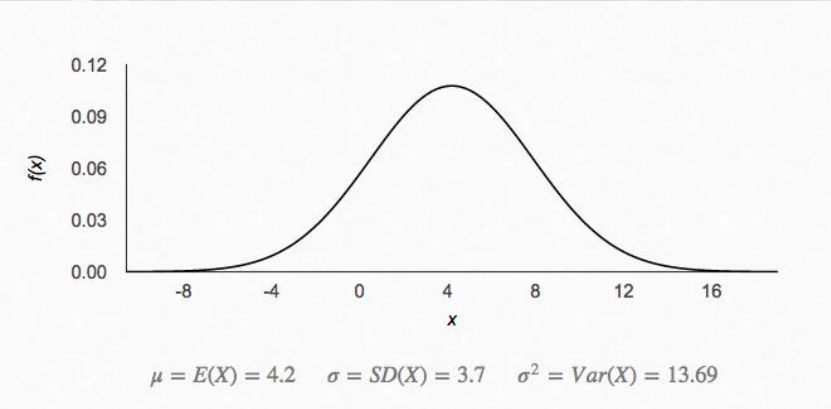
$N_{2,\mathbf{POS}} = N(x; 7.1, 2.0)$



$N_{1,\mathbf{NEG}} = N(x; 4.7, 2.5)$



$N_{2,\mathbf{NEG}} = N(x; 4.2, 3.7)$



Now, suppose you have a new example **x**, with $x_1 = 5.2$, $x_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

Now, suppose you have a new example **x**, with $x_1 = 5.2$, $x_2 = 6.3$.

What is $class_{NB}(\mathbf{x})$?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note: N is the probability density function, but can be used analogously to probability in Naïve Bayes calculations.

Now, suppose you have a new example **x**, with $x_1 = 5.2$, $x_2 = 6.3$.

What is $class_{NB}(\mathbf{x})$?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(x_1 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(1.8)} e^{-\frac{(5.2-4.8)^2}{2(1.8)^2}} = .22$$

$$P(x_2 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(2.0)} e^{-\frac{(6.3-7.1)^2}{2(2.0)^2}} = .18$$

$$P(x_1 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(2.5)} e^{-\frac{(5.2-4.7)^2}{2(2.5)^2}} = .16$$

$$P(x_2 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(3.7)} e^{-\frac{(6.3-4.2)^2}{2(3.7)^2}} = .09$$

Positive:

$$P(\mathbf{POS})P(x_1 \mid \mathbf{POS})P(x_2 \mid \mathbf{POS}) = (.5)(.22)(.18) = .02$$

Negative:

$$P(\mathbf{NEG})P(x_1 \mid \mathbf{NEG})P(x_2 \mid \mathbf{NEG}) = (.5)(.16)(.09) = .0072$$

$class_{NB}$ (**x**) = **POS**

Use logarithms to avoid underflow

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class)\prod_i P(x_i \mid class)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; \log\left(P(class)\prod_i P(x_i \mid class) \right)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \left(\log P(class) + \sum_i \log P(x_i \mid class) \right)$$

Learning conditional probabilities

- In general, random variables are not binary, but real-valued
- Conditional probability tables conditional probability distributions
- Estimate parameters of these distributions from data
- If data is missing on one or more variables, use "expectation maximization" algorithm

Approximate inference via sampling

- Recall: We can calculate full joint probability distribution from network.

$$P(X_1,\ldots,X_d)=\prod_{i=1}^{d}P(X_i \mid parents(X_i))$$

  where parents(X_i) denotes specific values of parents of X_i.

- We can do diagnostic, causal, and inter-causal inference

- But if there are a lot of nodes in the network, this can be very slow!

  Need efficient algorithms to do approximate calculations!

Applying Bayesian Reasoning to Speech Recognition

- **Task:** Identify sequence of words uttered by speaker, given acoustic signal.
- Uncertainty introduced by noise, speaker error, variation in pronunciation, homonyms, etc.
- Thus speech recognition is viewed as problem of probabilistic inference.

- So far, we've looked at probabilistic reasoning in static environments.

- Speech: Time sequence of "static environments".
  - Let **X** be the "state variables" (i.e., set of non-evidence variables) describing the environment (e.g., *Words* said during time step **t**)

  - Let **E** be the set of evidence variables (e.g., **S** = features of acoustic signal).

– The **E** values and **X** joint probability distribution changes over time.

t_1: $\mathbf{X}_1$, $\mathbf{e}_1$

t_2: $\mathbf{X}_2$, $\mathbf{e}_2$

etc.

- At each t, we want to compute $P(Words \mid \mathbf{S})$.

- We know from Bayes rule:

$$P(Words \mid \mathbf{S}) = \alpha\, P(\mathbf{S} \mid Words) P(Words)$$

- $P(\boldsymbol{S} \mid Words)$, for all words, is a previously learned "acoustic model".
  - E.g. For each word, probability distribution over phones, and for each phone, probability distribution over acoustic signals (which can vary in pitch, speed, volume).
- $\boldsymbol{P}(Words)$, for all words, is the "language model", which specifies prior probability of each utterance.

  - E.g. "**bigram model**": probability of each word following each other word.

- Speech recognition typically makes three assumptions:

1. Process underlying change is itself "stationary"
   i.e., state transition probabilities don't change

2. Current state **X** depends on only a finite history of previous states ("**Markov assumption"**).
   – Markov process of order *n*: Current state depends only on *n* previous states.

3. Values $\mathbf{e}_t$ of evidence variables depend only on current state $\mathbf{X}_t$. ("**Sensor model**")

Phones

All human speech is composed from 40-50 phones, determined by the configuration of articulators (lips, teeth, tongue, vocal cords, air flow)

Form an intermediate level of hidden states between words and signal

$\Rightarrow$ acoustic model = pronunciation model + phone model

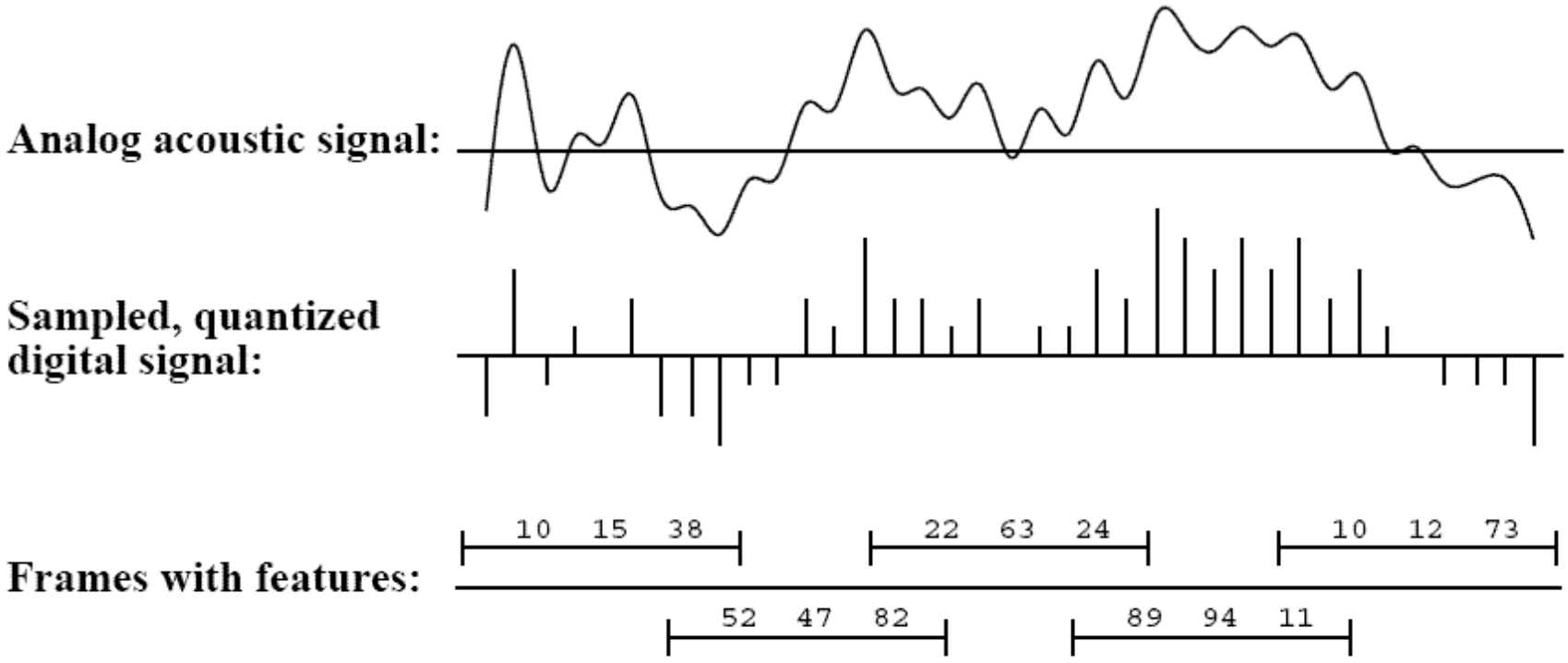
ARPAbet designed for American English

[iy]	b**ea**t	[b]	**b**et	[p]	**p**et
[ih]	b**i**t	[ch]	**Ch**et	[r]	**r**at
[ey]	b**e**t	[d]	**d**ebt	[s]	**s**et
[ao]	b**ough**t	[hh]	**h**at	[th]	**th**ick
[ow]	b**oa**t	[hv]	**h**igh	[dh]	**th**at
[er]	B**er**t	[l]	**l**et	[w]	**w**et
[ix]	ros**e**s	[ng]	si**ng**	[en]	butt**on**
⋮	⋮	⋮	⋮	⋮	⋮

E.g., "ceiling" is [s iy l ih ng] / [s iy l ix ng] / [s iy l en]

Speech sounds

Raw signal is the microphone displacement as a function of time; processed into overlapping 30ms frames, each described by features



Frame features are typically formants—peaks in the power spectrum

Hidden Markov Models

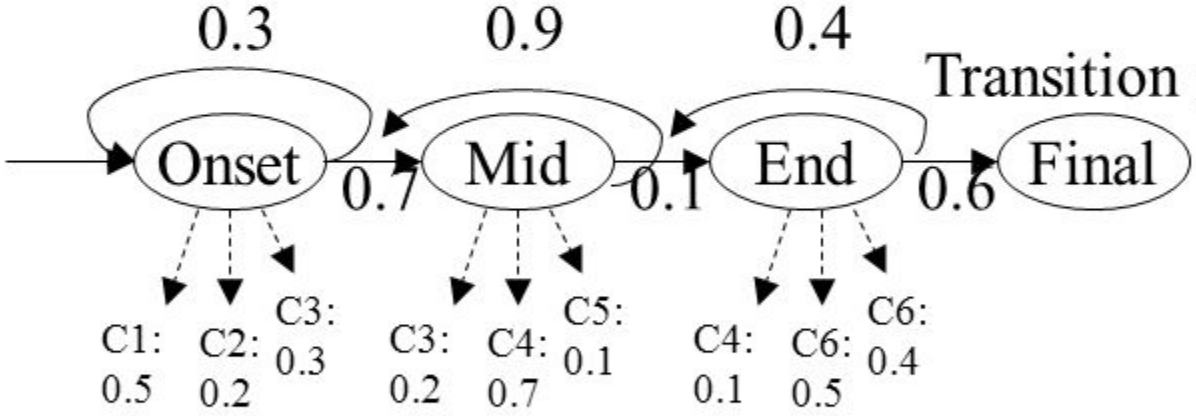
- **Markov model**: Given state $\boldsymbol{X}_t$, what is probability of transitioning to next state $\boldsymbol{X}_{t+1}$?
    - E.g., word bigram probabilities give

      $P\,(word_{t+1} \mid word_t\,)$

- **Hidden Markov model:** There are observable states (e.g., signal **S**) and "hidden" states (e.g., *Words*). **HMM** represents probabilities of hidden states given observable states.

Acoustic Model

- 3-state phone model for [m]
  - Use Hidden Markov Model (HMM)

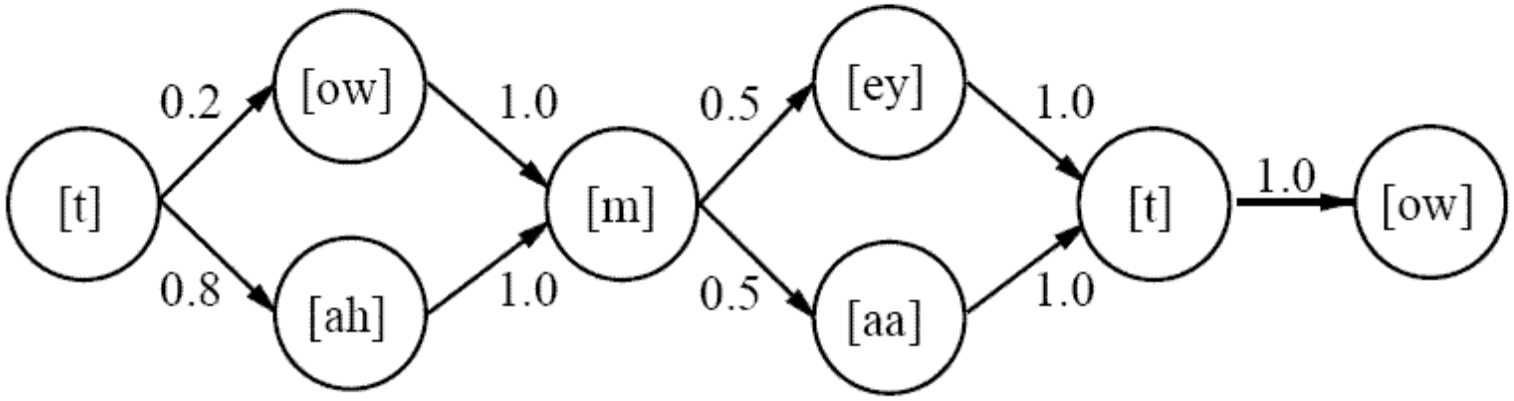


  - Probability of sequence: sum of prob of paths

Word pronunciation models

Each word is described as a distribution over phone sequences

Distribution represented as an HMM transition model



$$P([towmeytow]|\text{“tomato”}) = P([towmaatow]|\text{“tomato”}) = 0.1$$
$$P([tahmeytow]|\text{“tomato”}) = P([tahmaatow]|\text{“tomato”}) = 0.4$$

Structure is created manually, transition probabilities learned from data

Continuous speech

Not just a sequence of isolated-word recognition problems!

- Adjacent words highly correlated
- Sequence of most likely words $\neq$ most likely sequence of words
- Segmentation: there are few gaps in speech
- Cross-word coarticulation—e.g., "next thing"

Continuous speech systems manage 60–80% accuracy on a good day

Example: "I'm firsty, um, can I have something to dwink?"

Language model

Prior probability of a word sequence is given by chain rule:

$$P(w_1 \cdots w_n) = \prod_{i=1}^{n} P(w_i | w_1 \cdots w_{i-1})$$

Bigram model:

$$P(w_i | w_1 \cdots w_{i-1}) \approx P(w_i | w_{i-1})$$

Train by counting all word pairs in a large text corpus

More sophisticated models (trigrams, grammars, etc.) help a little bit