

ECE5984 – Applications of Machine Learning

Lecture 13 – Probability-based Learning

Creed Jones, PhD

Course update

- HW3 due TODAY by 11:59 PM
- Project I has been posted
 - Due Tuesday, March 22
- Quiz 4 this Thursday, March 3
 - Lectures 11, 12 and 13
- Spring break next week

A few homework questions

- Consistency of the dataset
- Creating and saving the graph using graphviz
- Recursive steps of the ID3 algorithm

Today's Objectives

A brief discussion of clustering

Probability-based Learning

- Bayes' Theorem
- Bayesian Prediction
- Conditional Independence and Factorization

What about K-means? Is that classification?

- No, but it's related
- K-means is generally used to impose K classes onto unlabeled data
 - Imagine a room full of dogs of unknown breeds; if we assume there are five breeds present, K-means can split them into the “best” five groups
 - In K -means, the K stands for the number of classes to assume
 - The trick is to know what K to use
- In machine learning, we usually have a training set of labeled samples, and want to assign one or more new samples (*queries*) to the best label
 - In k -nearest neighbor, the k does not stand for the number of classes, but for the number of neighbors to use in assigning the label to the query

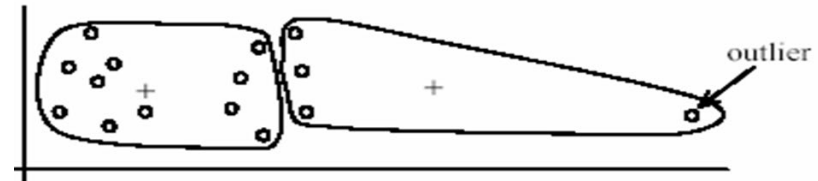
K-means: pros and cons

Pros

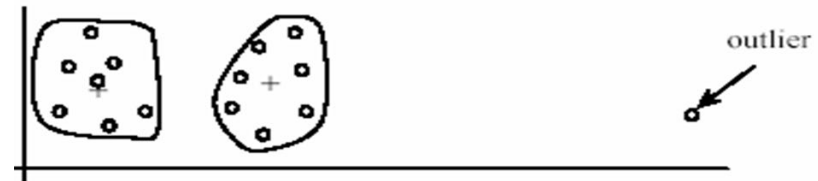
- Simple, fast to compute
- Converges to local minimum of within-cluster squared error

Cons/issues

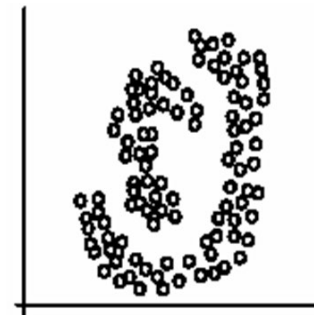
- Setting k ?
- Sensitive to initial centers
- Sensitive to outliers
- **Detects spherical clusters**
- Assuming means can be computed



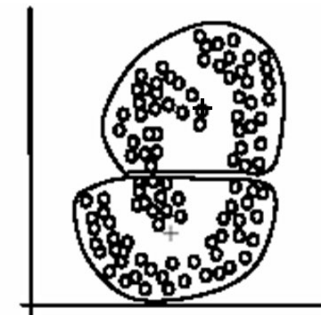
(A): Undesirable clusters



(B): Ideal clusters



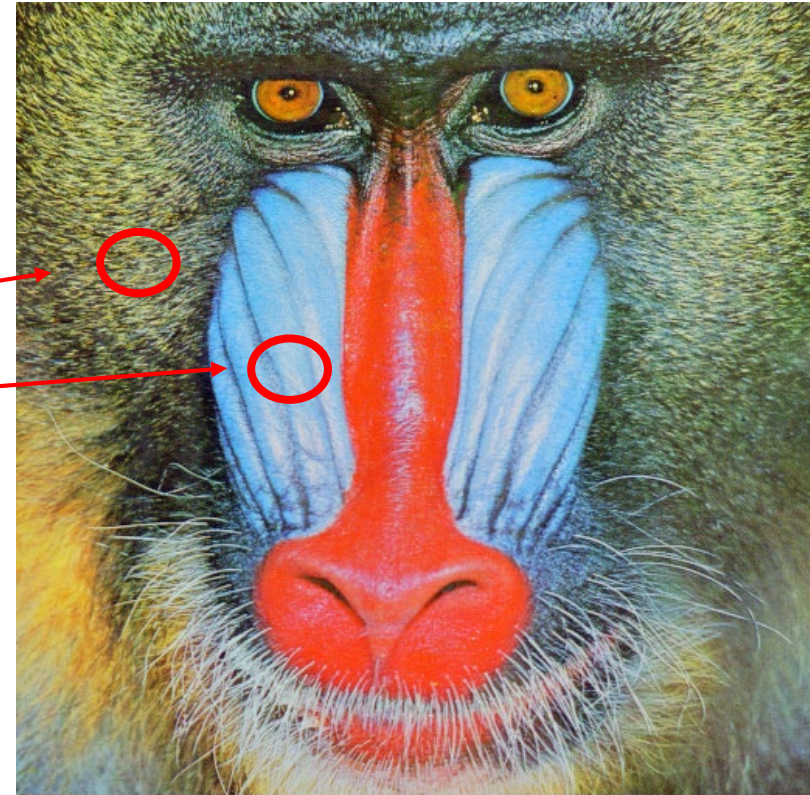
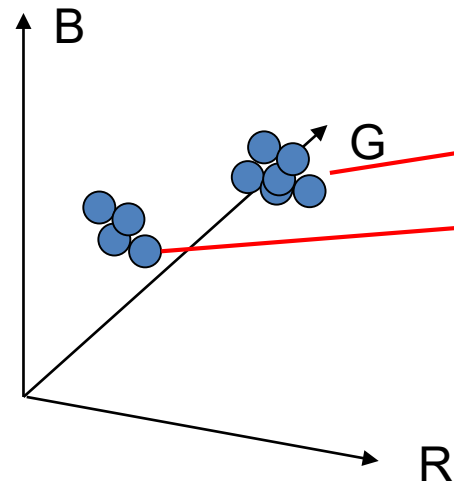
(A): Two natural clusters



(B): k -means clusters

Image Segmentation as clustering

Grouping pixels based on the color similarity



Look at the output of KMeans clustering
in RGB space on a color image

- find clusters
- set each pixel to its cluster number
- assign false colors to each cluster



Look at the output of KMeans clustering
in RGB space on a color image

- find clusters
- set each pixel to its cluster number
- assign false colors to each cluster



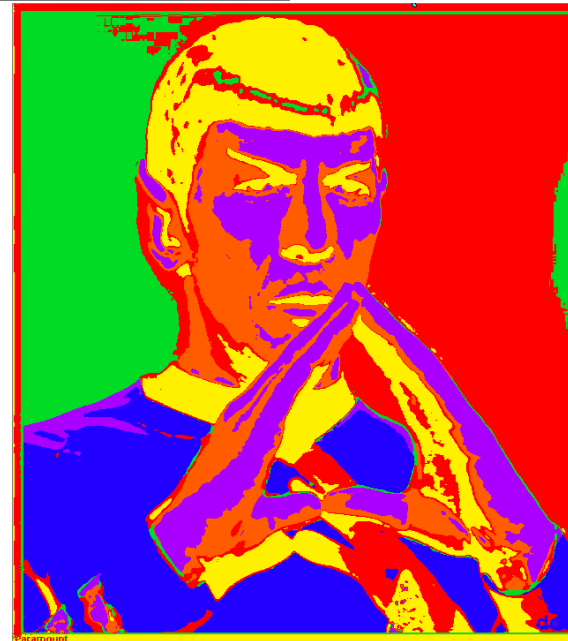
K=4



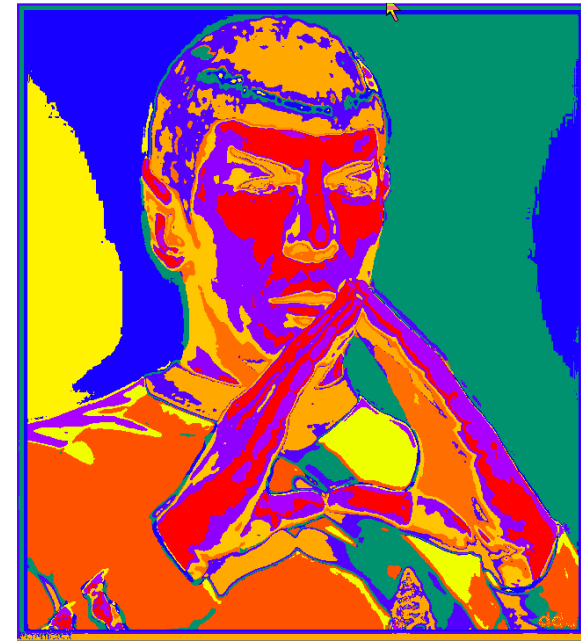
K=5



K=6



K=12



Color space Kmeans clustering in opencv/Python

```
import numpy as np
import cv2
import sklearn.cluster as cl

nclusters = 5
# Load an image in grayscale
img = cv2.imread('C:\\Data\\spock.jpg')
print(img.shape)
colors = img.reshape((img.shape[0]*img.shape[1], 3))
clus = cl.KMeans(nclusters)
clus.fit(colors)
newcolors = np.uint8(clus.predict(colors))
newimg = newcolors.reshape((img.shape[0], img.shape[1]))
pcolor = cv2.applyColorMap(cv2.equalizeHist(newimg), cv2.COLORMAP_RAINBOW)

cv2.imshow('INPUT',img)
cv2.imshow('NEW', pcolor)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

PROBABILITY-BASED LEARNING

This simple game involved guessing which of the three cards is the queen.



(a)

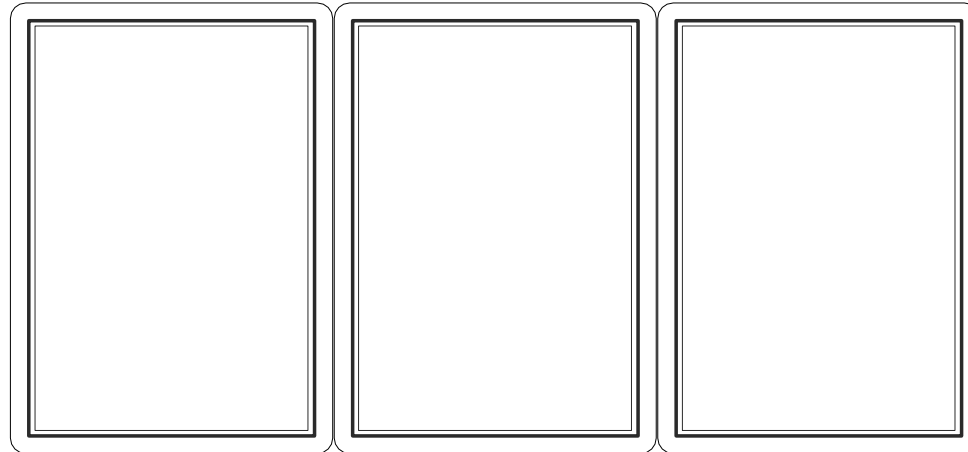
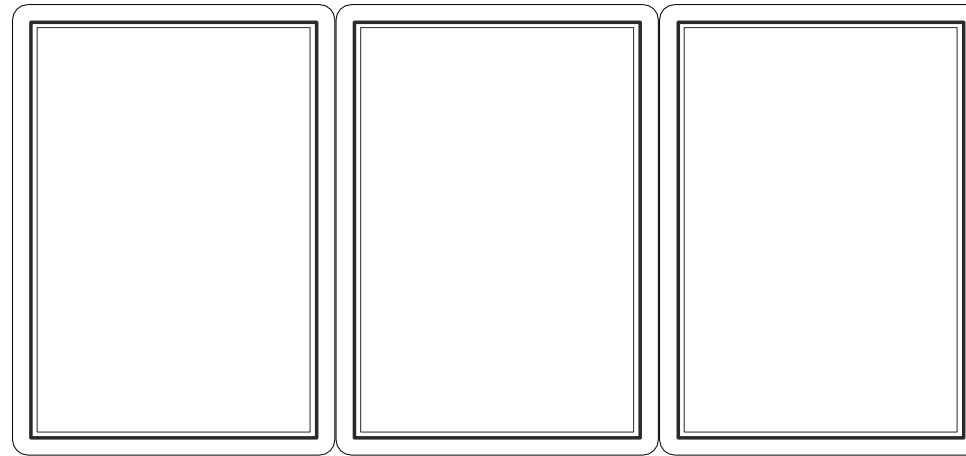


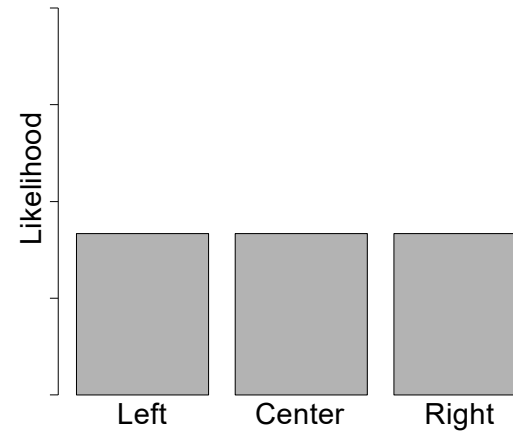
Figure: A game of *find the lady*

If we have no other information, then our best estimate of the probability that the queen is in each of the three spots is $1/3$.

The three positions are equiprobable.



(a)



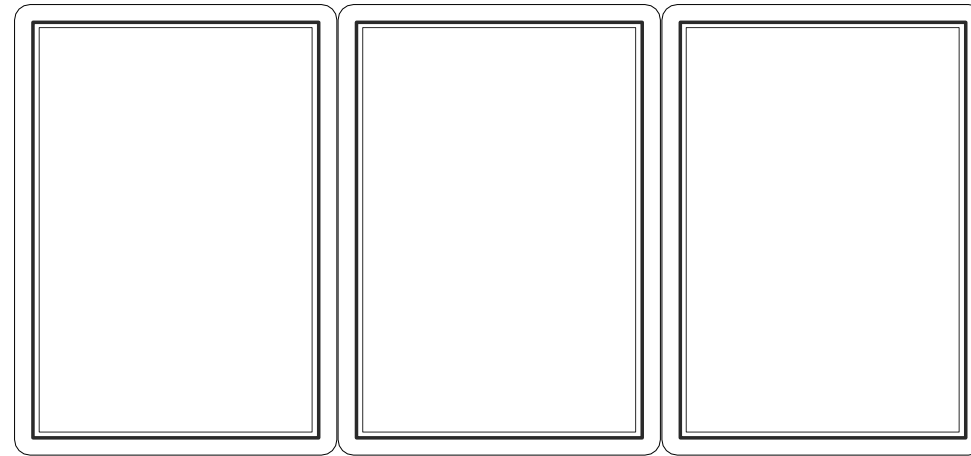
(b)

Figure: A game of *find the lady* : (a) the cards dealt face down on a table; and (b) the initial likelihoods of the queen ending up in each position.

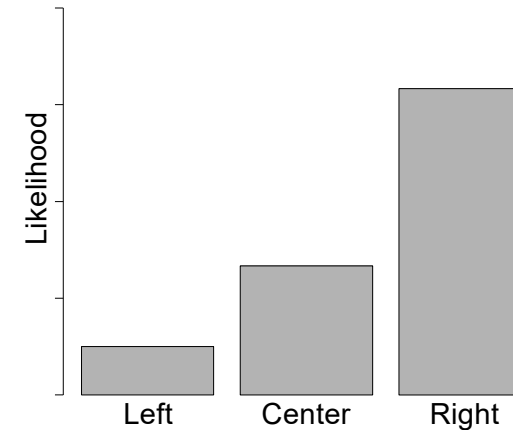
After watching the dealer play 30 games with other players, you notice that he has a tendency to drop the queen in the position on the right (19 times) more than the left (3 times) or center (8 times).

Based on this, you update you beliefs about where the queen is likely to land based on the evidence that you have collected:

- $p(left) = \frac{3}{30}$
- $p(center) = \frac{8}{30}$
- $p(right) = \frac{19}{30}$



(a)



(b)

Figure: A game of *find the lady* : (a) the cards dealt face down on a table; and (b) a revised set of likelihoods for the position of the queen based on evidence collected.

Now, a gust of wind flips the rightmost card over – and it’s an ace. So the probability that it’s a queen is now 0.

Remember the dealer’s tendency to drop the queen in the position on the right (19 times) more than the left (3 times) or center (8 times).

So:

- $p(left) = \frac{3}{11}$
- $p(center) = \frac{8}{11}$
- $p(right) = 0$

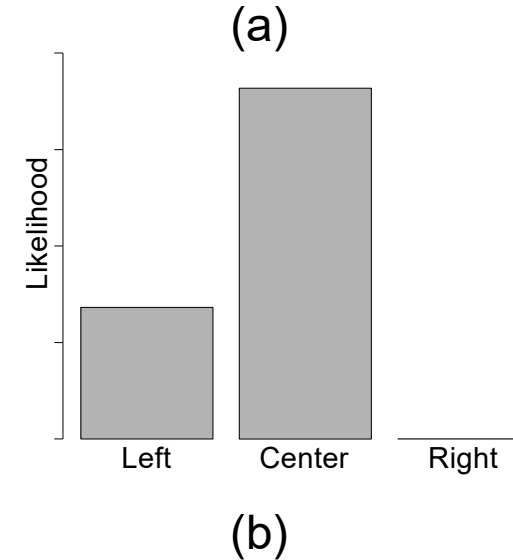
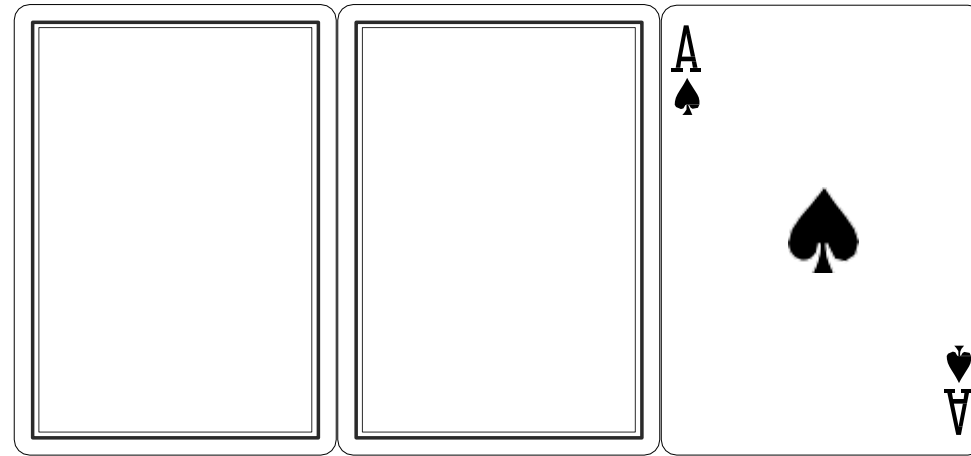


Figure: A game of *find the lady* : (a) The set of cards after the wind blows over the one on the right; (b) the revised likelihoods for the position of the queen based on this new evidence.

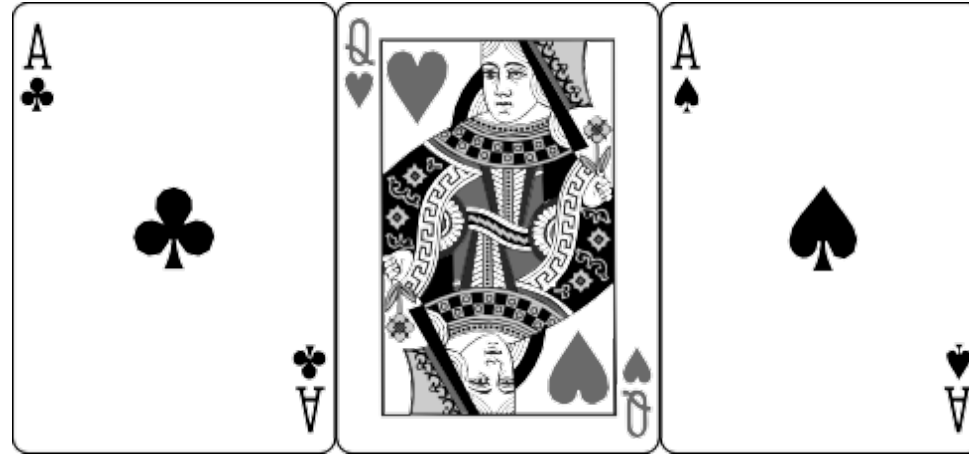


Figure: A game of *find the lady* : The final positions of the cards in the game.

Big Idea

- ♥ We can use estimates of likelihoods to determine the most likely prediction that should be made.
- ♥ More importantly, we revise these predictions based on data we collect and whenever extra evidence becomes available.

FUNDAMENTALS

A probability function $P(X = x)$ returns the probability of a feature X taking a specific value x

- A joint probability refers to the probability of an assignment of specific values to multiple different features
- A conditional probability refers to the probability of one feature taking a specific value given that we already know the value of a different feature
- A probability distribution is a data structure that describes the probability of each possible value a feature can take. The sum of a probability distribution must equal 1.0.
- A joint probability distribution is a probability distribution over more than one feature assignment and is written as a multi-dimensional matrix in which each cell lists the probability of a particular combination of feature values being assigned
- The sum of all the cells in a joint probability distribution must be 1.0

Examine the joint distribution calculated from a small dataset of meningitis diagnosis and symptoms

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

$$P(H, F, V, M) = \begin{bmatrix} P(h, f, v, m) & P(\bar{h}, f, v, m) \\ P(h, f, v, \bar{m}) & P(\bar{h}, f, v, \bar{m}) \\ P(h, f, \bar{v}, m) & P(\bar{h}, f, \bar{v}, m) \\ P(h, f, \bar{v}, \bar{m}) & P(\bar{h}, f, \bar{v}, \bar{m}) \\ P(h, \bar{f}, v, m) & P(\bar{h}, \bar{f}, v, m) \\ P(h, \bar{f}, v, \bar{m}) & P(\bar{h}, \bar{f}, v, \bar{m}) \\ P(h, \bar{f}, \bar{v}, m) & P(\bar{h}, \bar{f}, \bar{v}, m) \\ P(h, \bar{f}, \bar{v}, \bar{m}) & P(\bar{h}, \bar{f}, \bar{v}, \bar{m}) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0 \\ 0.4 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

We can compute the probability of any event by summing over the cells in the joint distribution where that event is true - this is called *summing out*

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

$$P(H, F, V, M) = \begin{bmatrix} P(h, f, v, m) & p(\bar{h}, f, v, m) \\ \mathbf{P(h, f, v, \bar{m})} & P(\bar{h}, f, v, \bar{m}) \\ P(h, f, \bar{v}, m) & P(\bar{h}, f, \bar{v}, m) \\ \mathbf{P(h, f, \bar{v}, \bar{m})} & P(\bar{h}, f, \bar{v}, \bar{m}) \\ P(h, \bar{f}, v, m) & P(\bar{h}, \bar{f}, v, m) \\ \mathbf{P(h, \bar{f}, v, \bar{m})} & P(\bar{h}, \bar{f}, v, \bar{m}) \\ P(h, \bar{f}, \bar{v}, m) & P(\bar{h}, \bar{f}, \bar{v}, m) \\ \mathbf{P(h, \bar{f}, \bar{v}, \bar{m})} & P(\bar{h}, \bar{f}, \bar{v}, \bar{m}) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \mathbf{0} & 0 \\ 0 & 0.1 \\ \mathbf{0.1} & 0.2 \\ 0.2 & 0 \\ \mathbf{0.4} & 0 \\ 0 & 0 \\ \mathbf{0} & 0 \end{bmatrix}$$

$$p(\text{headache and no meningitis}) = p(h, \bar{m}) = 0 + 0.1 + 0.4 + 0 = \mathbf{0.5}$$

Bayes' Theorem

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

Bayes' theorem relates *prior* (before we have additional knowledge) and *posterior* (after we know more) probabilities

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Prior Probability

Likelihood of the evidence 'E' if the Hypothesis 'H' is true

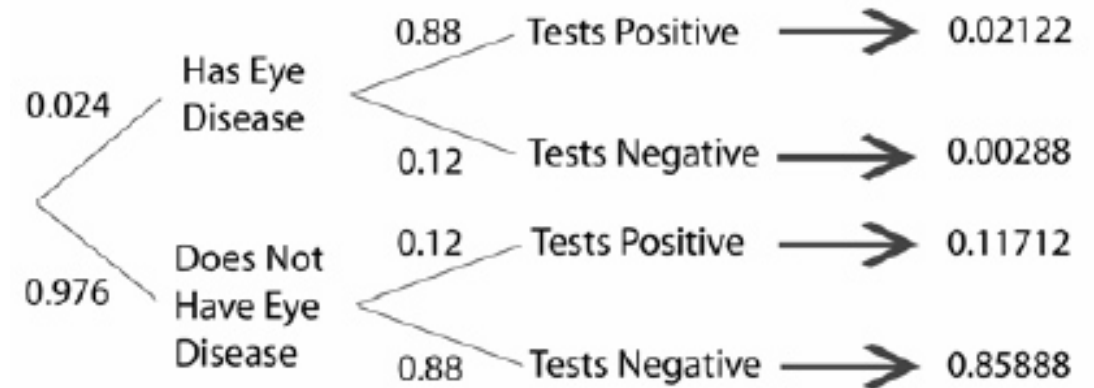
Posterior Probability of 'H' given the evidence

Priori probability that the evidence itself is true

The divisor is the prior probability of the evidence
This division functions as a normalization constant

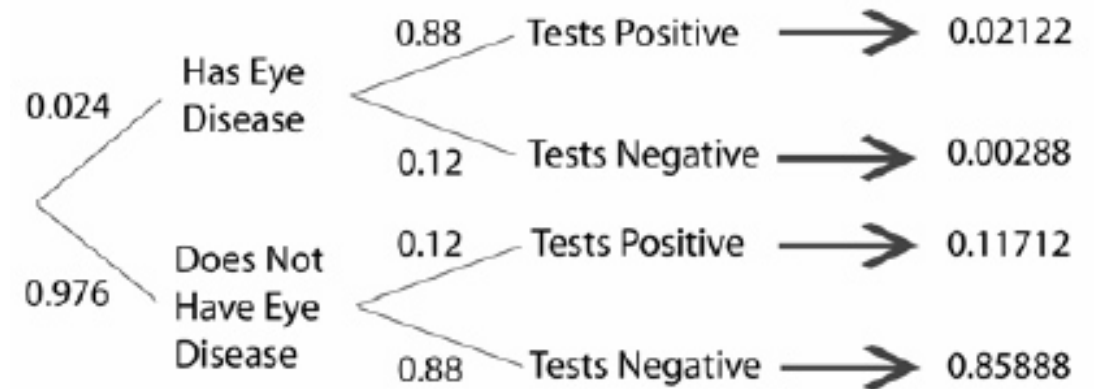
Bayes Theorem for Diagnostic Test

- This tree shows the probability of a positive test, given that the patient has the disease
- We are instead interested in the probability that the person has the disease, given a positive test.
- $p(test) = \frac{0.11712 + 0.02122}{1.000} = 0.13834$
- $p(dis|test) = \frac{p(dis)p(test|dis)}{p(test)} = \frac{0.024 \cdot 0.88}{0.13834} = 0.152$
- So those who fail this diagnostic still only have a 15% chance of having the eye disease!
- Think of it this way – 85% of the time that people are told “your test results show a possible problem”, it’s a false alarm!



Bayes Theorem for Diagnostic Test

- Now consider the probability that someone who does not fail the test actually has the disease:
- $$p(dis|\sim test) = \frac{p(dis)p(\sim test|dis)}{p(\sim test)} = \frac{0.024 \cdot 0.12}{1 - 0.13834} = 0.00334$$
- There is an extremely low possibility that those who actually suffer from the disease will go undetected
 - This is a really good diagnostic test
 - Very low probability of missing a positive case
 - Even with the high false alarm rate



USING BAYES' THEOREM

Bayes Theorem is Fundamental

$$p(B|A) = \frac{p(B)p(A|B)}{p(A)}$$

The likelihood of B occurring, given new evidence (that A occurred), can be calculated

It's equal to the original likelihood of B occurring...

...multiplied by the ratio of $p(A|B)$ to $p(A)$.
 This will be big (and increase $p(B|A)$) when most of the time A occurs is when B occurs.
 This will be small (and decrease $p(B|A)$) when A's likelihood is decreased by B occurring.

Example – Is auto travel more dangerous than air travel?

- I often hear people say "you're more likely to die in an auto accident on the way to the airport than to die in a plane crash."
 - Not sure if this is supposed to be reassuring...
- Is this true? Can we figure it out?
- Thesis: $p(\text{die in one car trip}) > p(\text{die in one air trip})$



We can estimate the relative probabilities, based on past events, to see if the thesis is at all realistic

$$p(\text{die in one car trip}) \stackrel{?}{>} p(\text{die in one air trip})$$

$$\frac{\text{AutoTrips resulting in Death}}{\text{AutoTrips}} \stackrel{?}{>} \frac{\text{Flights resulting in Death}}{\text{Flights}}$$

or maybe:

$$\frac{\text{Highway Fatalities}}{\text{AutoTrips}} \stackrel{?}{>} \frac{\text{Air Fatalities}}{\text{Flights}}$$

Let's gather some statistics on these events

$$\frac{\text{Highway Fatalities}}{\text{AutoTrips}} \stackrel{?}{>} \frac{\text{Air Fatalities}}{\text{Flights}}$$

- 28,537 commercial flights per day
 - <http://sos.noaa.gov/Datasets/dataset.php?id=44>
- 26 fatalities, 2004-2013
 - <http://www.boeing.com/news/techissues/pdf/statsum.pdf> , page 15
- 32,367 highway fatalities in 2011
 - <http://www.nhtsa.gov/NCSA>
- 3.79 car trips per person per day in 2009
 - <http://nhts.ornl.gov/2009/pub/stt.pdf>

Look at the probability of a highway fatality...

- 32,367 US highway fatalities in 2011
- 3.79 car trips per person per day in 2009



$$\frac{\text{Highway Fatalities}}{\text{AutoTrips}}$$

$$= \frac{32,637}{3.79(365)(316,000,000)} = 7.47 \times 10^{-8}$$

- So any particular auto trip has a 0.000000075 chance of killing you
 - Does this seem right?

Look at the probability of an airline fatality...

- 28,537 commercial flights per day
- 26 fatalities, 2004-2013



$$\frac{\textit{AirlineFatalities}}{\textit{AirTrips}}$$

$$= \frac{26}{28,537(10)(365)} = 2.5 \times 10^{-7}$$

- So any particular flight has a 0.00000025 chance of killing you

In conclusion, you are about three times more likely to die in any given flight than in any given auto trip

- $p(\text{DeathInOneFlight}) = 2.5 \times 10^{-7}$
- $p(\text{DeathInOneDrive}) = 7.5 \times 10^{-8}$
- The probability of death in a flight is still about $\frac{2.5 \times 10^{-7}}{7.5 \times 10^{-8}} \approx 3$ times higher than death in a drive
- However, in my mind, this still proves the point of just how safe air travel has become



Can we calculate the likelihood of dying, given that we are speeding? Apply Bayes' Rule to what we know

- $p(\text{dieInAutoAcc} \mid \text{speeding}) = \frac{p(\text{dieInAutoAcc})p(\text{speeding} \mid \text{dieInAutoAcc})}{p(\text{speeding})}$
- $p(\text{dieInAutoAcc}) = p(\text{DeathInOneDrive}) = 7.5 \times 10^{-8}$
- Assume 11% of drivers speed: $p(\text{speeding}) = 0.11$
 - “As many as 16% of male drivers and 6% of women motorists said they frequently exceeded limits”, according to insurance industry studies
 - <http://www.belfasttelegraph.co.uk/life/motoring/many-admit-driving-over-speed-limit-28784278.html>
- $p(\text{speeding} \mid \text{dieInAutoAcc}) = 0.29$
 - “In 2013, speeding was a factor in 29 percent of motor vehicle crash deaths”
 - <http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/overview-of-fatality-facts>
- So, $p(\text{dieInAutoAcc} \mid \text{speeding}) = \frac{7.5 \times 10^{-8}(0.29)}{0.11} = 1.977 \times 10^{-7}$
- Much closer to the $p(\text{DeathInOneFlight}) = 2.5 \times 10^{-7}$

We can derive Bayes' theorem from the basic definitions of conditional probability

The *conditional probability rule* states that:

$$p(A \text{ and } B) = p(A)p(B|A)$$

but this must also equal:

$$p(A \text{ and } B) = p(B)p(A|B)$$

So:

$$p(X|Y)p(Y) = p(Y|X)p(X)$$

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

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

Let's look at another example

Example

After a yearly checkup, a doctor informs their patient that he has both bad news and good news. The bad news is that the patient has tested positive for a serious disease and that the test that the doctor has used is 99% accurate (i.e., the probability of testing positive when a patient has the disease is 0.99, as is the probability of testing negative when a patient does not have the disease). The good news, however, is that the disease is extremely rare, striking only 1 in 10,000 people.

What is the actual probability that the patient has the disease?

Why is the rarity of the disease good news given that the patient has tested positive for it?

Example

After a yearly checkup, a doctor informs their patient that he has both bad news and good news. The bad news is that the patient has tested positive for a serious disease and that the test that the doctor has used is 99% accurate (i.e., the probability of testing positive when a patient has the disease is 0.99, as is the probability of testing negative when a patient does not have the disease). The good news, however, is that the disease is extremely rare, striking only 1 in 10,000 people.

- $p(d) = 0.0001$, $p(t|d) = 0.99$, $p(\bar{t}|\bar{d}) = 0.99$, therefore $p(t|\bar{d}) = 0.01$
- $p(disease|test) = p(d|t) = \frac{p(t|d)p(d)}{p(t)}$
- $p(t) = p(t|d)p(d) + p(t|\bar{d})p(\bar{d}) = (0.99 \cdot 0.0001) + (0.01 \cdot 0.9999) = 0.0101$
- $p(d|t) = \frac{0.99 \cdot 0.0001}{0.0101} = 0.0098$

Bayes' rule, $p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$, computes an adjusted probability, using the *prior probability* of the related event

- The divisor is the prior probability of the evidence; this division functions as a normalization constant
- We can calculate this divisor directly from the dataset.

$$p(Y) = \frac{\# \text{ rows where } Y \text{ is true}}{\# \text{ rows}}$$

- Or, we can use the **Theorem of Total Probability** to calculate this divisor.

$$p(Y) = \sum_i p(Y|X_i)p(X_i)$$

The generalized form of Bayes' theorem relates to cases with several prior events

- If many events form the prior, then we use the joint distribution to calculate the probability of all of them occurring
- The form below relates the likelihood of the target variable t taking on the value l , given that events $q[1], q[2] \cdots q[m]$ have all occurred

Generalized Bayes' Theorem

$$P(t = l | \mathbf{q}[1], \dots, \mathbf{q}[m]) = \frac{P(\mathbf{q}[1], \dots, \mathbf{q}[m] | t = l) P(t = l)}{P(\mathbf{q}[1], \dots, \mathbf{q}[m])}$$

Chain Rule

$$P(\mathbf{q}[1], \dots, \mathbf{q}[m]) = \\ P(\mathbf{q}[1]) \times P(\mathbf{q}[2]|\mathbf{q}[1]) \times \\ \dots \times P(\mathbf{q}[m]|\mathbf{q}[m-1], \dots, \mathbf{q}[2], \mathbf{q}[1])$$

To apply the chain rule to a conditional probability we just add the conditioning term to each term in the expression:

$$P(q[1], \dots, q[m] | t = l) = \\ P(q[1] | t = l) \times P(q[2] | q[1], t = l) \times \dots \\ \dots \times P(q[m] | q[m-1], \dots, q[3], q[2], q[1], t = l)$$

Let's use Bayes rule to determine the probability of meningitis for a specific set of symptoms

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
q	true	false	true	?


- This is the data set presented earlier
- Note that the specific set of symptoms matches several entries
 - The data set is not *consistent*
- How can we calculate the probability of meningitis?

We want the probability of meningitis for these symptoms – what is $p(M|\{h, \bar{f}, v\})$?

- According to Bayes, $p(M|\{h, \bar{f}, v\}) = \frac{p(\{h, \bar{f}, v\}|M)p(M)}{p(\{h, \bar{f}, v\})}$
- From the data, $p(M) = \frac{3}{10} = 0.3$, $p(\{h, \bar{f}, v\}) = \frac{6}{10} = 0.6$
- $p(\{h, \bar{f}, v\}|M) = \frac{|\{d_8, d_{10}\}|}{|\{d_5, d_8, d_{10}\}|} = \frac{2}{3} = 0.6667$
- $p(M|\{h, \bar{f}, v\}) = \frac{p(\{h, \bar{f}, v\}|M)p(M)}{p(\{h, \bar{f}, v\})} = \frac{0.6667 \cdot 0.3}{0.6} = 0.3333$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

The Paradox of the False Positive

-  The mistake of forgetting to factor in the prior gives rise to the **paradox of the false positive** which states that in order to make predictions about a rare event the model has to be as accurate as the prior of the event is rare or there is a significant chance of **false positives** predictions (i.e., predicting the event when it is not the case).

Higher dimensions means fewer instances of each combination of feature values

Curse of Dimensionality

As the number of descriptive features grows the number of potential conditioning events grows. Consequently, an exponential increase is required in the size of the dataset as each new descriptive feature is added to ensure that for any conditional probability there are enough instances in the training dataset matching the conditions so that the resulting probability is reasonable.

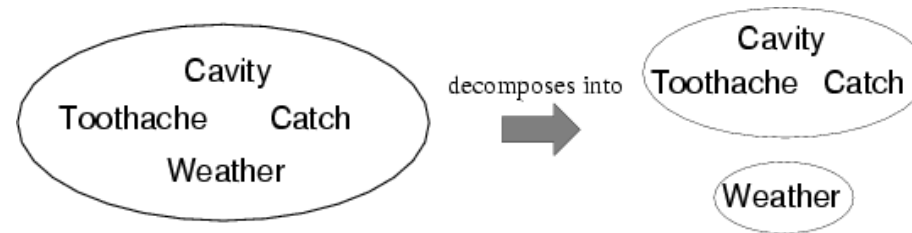
The probability of a patient who has a headache and a fever having meningitis should be greater than zero!

Our dataset is not large enough → our model is over-fitting to the training data. The concepts of conditional independence and factorization can help us overcome this flaw of our current approach.

CONDITIONAL INDEPENDENCE

Independence means that two or more events don't influence the probability of each other

- A and B are independent if and only if
 $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)P(B)$



$$P(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather}) = P(\text{Toothache}, \text{Catch}, \text{Cavity}) P(\text{Weather})$$

-
- Absolute independence is powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence means that A is independent of B – *if C is true*

- $P(\text{Toothache}, \text{Cavity}, \text{Catch})$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
 - (1) $P(\text{catch} | \text{toothache}, \text{cavity}) = P(\text{catch} | \text{cavity})$
- The same independence holds if I haven't got a cavity:
 - (2) $P(\text{catch} | \text{toothache}, \neg \text{cavity}) = P(\text{catch} | \neg \text{cavity})$
- *Catch* is **conditionally independent** of *Toothache* given *Cavity*:
 - $P(\text{Catch} | \text{Toothache}, \text{Cavity}) = P(\text{Catch} | \text{Cavity})$
- Equivalent statements:
 - $P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$
 - $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity})$

Conditional independence continued

- Write out full joint distribution using chain rule:

$$\begin{aligned}
 P(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) &= P(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) P(\textit{Catch}, \textit{Cavity}) \\
 &= P(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) P(\textit{Catch} \mid \textit{Cavity}) P(\textit{Cavity}) \\
 &= P(\textit{Toothache} \mid \textit{Cavity}) P(\textit{Catch} \mid \textit{Cavity}) P(\textit{Cavity})
 \end{aligned}$$

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n .
 Conditional independence is our most basic and robust form of knowledge about uncertain environments

Assuming conditional independence simplifies the calculation of the prior probability

Without conditional independence

$$P(X, Y, Z | W) = P(X | W) \times P(Y | X, W) \times P(Z | Y, X, W) \times P(W)$$

With conditional independence

$$P(X, Y, Z | W) = \underbrace{P(X | W)}_{\text{Factor 1}} \times \underbrace{P(Y, W)}_{\text{Factor 2}} \times \underbrace{P(Z, W)}_{\text{Factor 3}} \times \underbrace{P(W)}_{\text{Factor 4}}$$

Conditional Independence

For two events, X and Y , that are conditionally independent given knowledge of a third events, here Z , the definition of the probability of a joint event and conditional probability are:

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

$$P(X, Y|Z) = P(X|Z) \times P(Y|Z)$$

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

$$\begin{aligned} P(X, Y) &= P(X|Y) \times P(Y) \\ &= P(Y|X) \times P(X) \end{aligned}$$

X and Y are **dependent**

$$P(X|Y) = P(X)$$

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

X and Y are **independent**

Let's assume that headache, nausea and fever are independent of each other for patients with meningitis

Assuming the descriptive features are conditionally independent of each other given MENINGITIS we only need to store four factors:

- Factor1 : $\langle p(M) \rangle$
- Factor2 : $\langle p(h|M), p(h|\bar{M}) \rangle$
- Factor3 : $\langle p(f|M), p(f|\bar{M}) \rangle$
- Factor4 : $\langle p(v|M), p(v|\bar{M}) \rangle$

$$p(h, f, v, M) = p(M) \cdot p(h|M) \cdot p(f|M) \cdot p(v|M)$$

Calculate the factors from the data:

- Factor1 : $p(M) = \frac{3}{10} = 0.3$
- Factor2 : $p(h|M) = \frac{2}{3} = 0.6667$, $p(h|\bar{M}) = \frac{5}{7} = 0.7143$
- Factor3 : $p(f|M) = \frac{1}{3} = 0.3333$, $p(f|\bar{M}) = \frac{3}{7} = 0.4286$
- Factor4 : $p(v|M) = \frac{2}{3} = 0.6667$, $p(v|\bar{M}) = \frac{4}{7} = 0.5714$
- Let's calculate the probability of meningitis for a patient with headache and fever but no vomiting

$$p(M|\{h, f, \bar{v}\}) = \frac{p(h|M) \cdot p(f|M) \cdot p(\bar{v}|M) \cdot p(M)}{\sum_i p(h|M_i) \cdot p(f|M_i) \cdot p(\bar{v}|M_i) \cdot p(M_i)}$$

$$= \frac{0.6667 \cdot 0.3333 \cdot 0.3333 \cdot 0.3}{0.6667 \cdot 0.3333 \cdot 0.3333 \cdot 0.3 + 0.7143 \cdot 0.4286 \cdot 0.4286 \cdot 0.7}$$

$$= 0.1948$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Today's Objectives

A brief discussion of clustering

Probability-based Learning

- Bayes' Theorem
- Bayesian Prediction
- Conditional Independence and Factorization