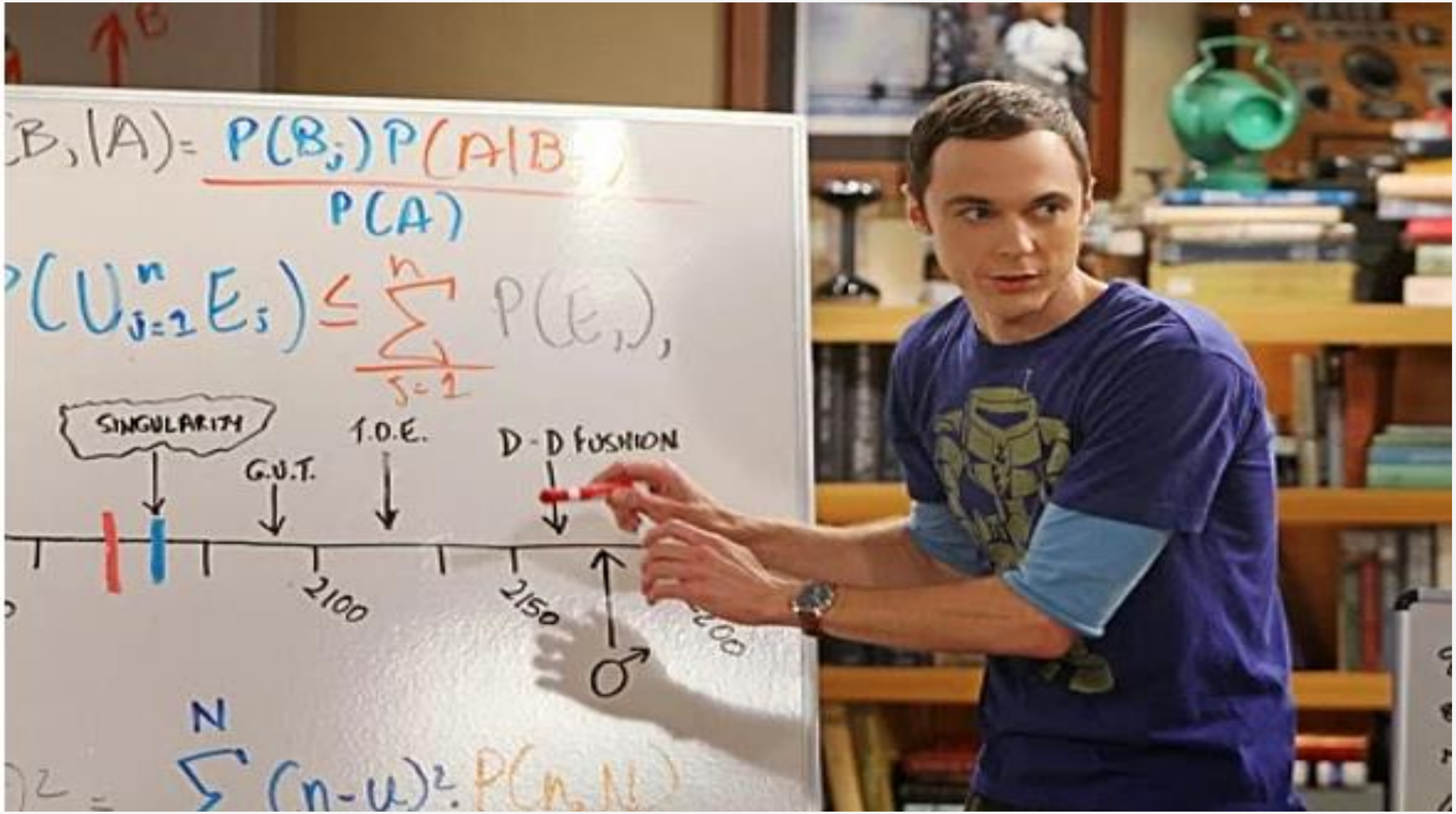
Sheldon Cooper is our guest lecturer today



Conditional Probability

The ***conditional probability*** of an event *B* is the probability that the event will occur given the knowledge that an event *A* has already occurred. This probability is written as *P(B|A), referred as Probability of B given A*

If events A and B are not independent, then the probability of the intersection of A and B (the probability that both events occur) is defined by

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

Or

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

Bayes Theorem

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

But A and B can be flipped: P(A and B) = P (B and A) = P(A|B) P (B), therefore

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

This is called the "Bayes Theorem".

Note: P (A) = P (A | B) P (B) + P (A | ~ B) P (~B) where the ~ B means the NOT B event (or complement of B)

Bayes Theorem

Problem:

- 1% of women have breast cancer (and therefore 99% do not).
- 80% of mammograms detect breast cancer when it is there (and therefore 20% miss it, i.e. 20% false negative).
- 10.0% of mammograms detect breast cancer when it's **not** there (and therefore 90% correctly return a negative result and 10% false positive).

What is the probability that the woman does have cancer if she is tested positive?

Solution:

Let A = Test Positive, B = Has Cancer, so we want to calculate $P(B \mid A)$

Given $P(B) = 1\%$, $P(A \mid B) = 80\%$, $P(A \mid \sim B) = 10\%$

Bayes Theorem

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A|B)\,P(B) + P(A|\sim B)\,P(\sim B)}$$

$$P(\text{Cancer}|\text{Positive}) = \frac{P(\text{Positive}|\text{Cancer})\,P(\text{Cancer})}{P(\text{Positive}|\text{Cancer})\,P(\text{Cancer}) + P(\text{Positive}|\text{No Cancer})\,P(\text{No Cancer})}$$

$$= \frac{80\% * 1\%}{80\% * 1\% + 10\% * 99\%} = 7.5\%$$

Bayes Theorem

- Interesting — a positive mammogram only means you have a 7. 5% chance of cancer, rather than 80% (the supposed accuracy of the test). It might seem strange at first but it makes sense: the test gives a false positive 10% of the time (quite high), so there will be many false positives in a given population. For a rare disease, most of the positive test results will be wrong.

- Let's test our intuition by drawing a conclusion from simply eyeballing the table. If you take 100 people, only 1 person will have cancer (1%), and they're most likely going to test positive (80% chance). Of the 99 remaining people, about 10% will test positive, so we'll get roughly 10 false positives. Considering all the positive tests, just 1 in 11 is correct, so there's a 1/11 chance of having cancer given a positive test. The real number is 7.5% (closer to 1/13, computed above), but we found a reasonable estimate without a calculator.

Naïve Bayes Classifier

How could we use Bayes theorem in Machine Learning?

The idea is in supervised learning, we know P(data | class label) from the training set,
What we need in predicting new data is in fact P(class label | data).

So we can use Bayes theorem:

$$P(\text{ class label on given observed data}) = \frac{P(\text{ observed data for each class})}{P(\text{ observed data })}$$

Bayesian Spam Filtering

- Example: Classify whether an email is Spam or not

- Class label: Spam or Not Spam (Ham)
- Data: Words inside the message

$$P(\text{ spam } | \text{ words }) = \frac{P(\text{ words } | \text{ spam })\, P(\text{spam})}{P(\text{words})}$$

- Will come back to this after we cover Natural Language Processing

- Let’s get back to the general Naïve Bayes Classifier

Mathematics behind Naïve Bayes Classifier

$$P(\text{label} \mid \text{features}) = \frac{P(\text{features} \mid \text{label})\, P(\text{label})}{P(\text{features})}$$

Remember we usually denote the label by y, the features are x_1, x_2, …,x_n,
We have

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)}$$

Make the naïve assumption that the features random variables are independent, i.e.

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y)$$

Mathematics behind Naïve Bayes Classifier

We have

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

Since P(x1, …, x_n) is constant given the input dataset, we can use the classification rules

$$P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

That is, we predict y by choosing the class that maximize the A Posteriori (MAP) distribution (P (y | x))

Mathematics behind Naïve Bayes Classifier

P(y) is just the relative frequency of the class label y. Harder question now is how to compute
P (x_i | y) ?

There are different naïve Bayes classifier that differ mainly by the assumptions they make regarding the distribution of P(x_i | y)

One common choice for P (x_i | y) is to assume that it is a Gaussian Distribution, which is parametrized by a mean (mu_c) and standard deviation (sigma_c), both of which can be estimated from the data

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Naïve Bayes Classifier

Learning by doing