Review

- We have provided a basic review of the probability theory
  - What is a (**discrete**) random variable
  - Basic axioms and theorems
  - Conditional distribution
  - Bayes rule

Bayes Rule

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

More general forms:

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

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

Commonly used discrete distributions

Bernoulli distribution: Ber(p)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = p^x (1-p)^{1-x}$$

Binomial distribution: x ~ Binomial(n , p)

the probability to see x heads out of n flips

$$\boxed{P(x) = \frac{n(n-1)\cdots(n-x+1)}{x!} p^x (1-p)^{n-x}}$$

Categorical distribution: x can take K values, the distribution is specified by a set of θ_k's

θ_k=P(x=v_k), and $\theta_1 + \theta_2 + \ldots + \theta_K = 1$

Multinomial distribution: Multinomial (n , [x_1, x_2, …, x_k])

The probability to see x_1 ones, x_2 twos, etc, out of n dice rolls

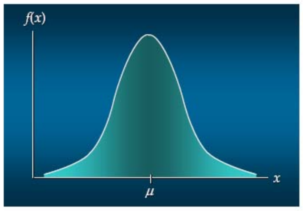
$$\boxed{P([x_1, x_2, \ldots, x_k]) = \frac{n!}{x_1! x_2! \cdots x_k!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}}$$

Continuous Probability Distribution

- A continuous random variable x can take any value in an interval on the real line
  - X usually corresponds to some real-valued measurements, e.g., today's lowest temperature
  - It is not possible to talk about the probability of a continuous random variable taking an exact value --- P(x=56.2)=0
  - Instead we talk about the probability of the random variable taking a value within a given interval P(x∈[50, 60])
  - This is captured in ***Probability density function***

PDF: probability density function

- The probability of X taking value in a given range [x1, x2] is defined to be the area under the PDF curve between x1 and x2
- We use *f*(x) to represent the PDF of x
- Note:
  - *f*(x) ≥0
  - *f*(x) can be larger than 1
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $P(X \in [x1, x2]) = \int_{x1}^{x2} f(x)dx$



What is the intuitive meaning of *f(x)*?

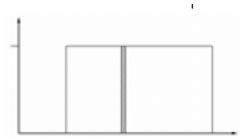
If *f* (x1)=α*a and *f* (x2)=a

Then when x is sampled from this distribution, you are α times more likely to see that x is "very close to" x1 than that x is "very close to" x2

Commonly Used Continuous Distributions

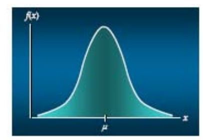
Uniform Probability Density Function

$$f(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



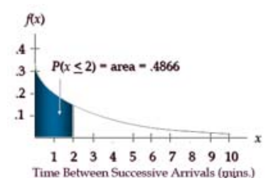
Normal (Gaussian) Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Exponential Probability Distribution

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$



- So far we have looked at univariate distributions, i.e., single random variables
- Now we will briefly look at joint distribution of multiple variables
- Why do we need to look at joint distribution?
  - Because sometimes different random variables are clearly related to each other
- Imagine three random variables
  - A: teacher appears grouchy
  - B: teacher had morning coffee
  - C: kelly parking lot is full at 8:50 AM
- How do we represent the distribution of 3 random variables together?

The Joint Distribution

Example: Binary variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Binary variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

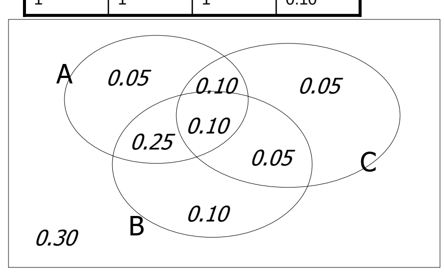
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Question: What is the relationship between p(A,B,C) and p(A)?

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

So we have learned that

- Joint distribution is useful!
  we can do all kinds of cool inference
  - I've got a sore neck: how likely am I to have meningitis?
  - Many industries grow around this kind of Inference: examples include medicine, pharma, Engine diagnosis etc.
- But, **HOW** do we get joint distribution?
  - We can learn from data

So we have learned that

- Joint distribution is extremely useful!
  we can do all kinds of cool inference
  - I've got a sore neck: how likely am I to have meningitis?
  - Many industries grow around Beyesian Inference: examples include medicine, pharma, Engine diagnosis etc.
- But, **HOW** do we get joint distribution?
  - We can learn from data

Learning a joint distribution

Build a JD table for your attributes in which the probabilities are unspecified

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	**0.25**
1	1	1	0.10

Fraction of all records in which A and B are True but C is False

Example of Learning a Joint

- This Joint was obtained by learning from three attributes in the UCI "Adult" Census Database [Kohavi 1995]

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

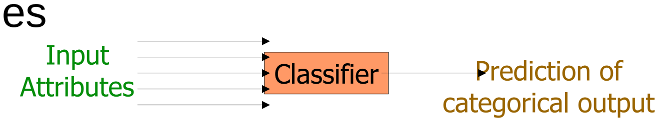
UCI machine learning repository:
http://www.ics.uci.edu/~mlearn/MLRepository.html

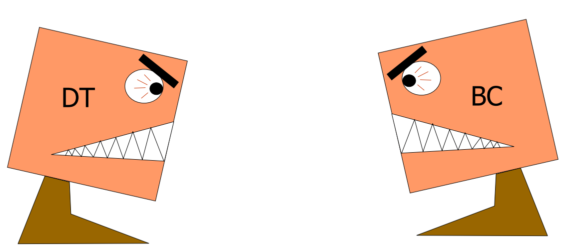
Where are we?

- We have recalled the fundamentals of probability
- We have become content with what JDs are and how to use them
- And we even know how to learn JDs from data.

Bayes Classifiers

- A formidable and sworn enemy of decision trees





Recipe for a Bayes Classifier

- Assume you want to predict output *Y* which has arity n_Y and values v_1, v_2, … v_{ny}.
- Assume there are *m* input attributes called X=(X_1, X_2, … X_m)
- Learn a conditional distribution of p(X|y) for each possible y value, y = v_1, v_2, … $v_{ny,}$, we do this by:
  - Break training set into n_Y subsets called DS_1, DS_2, … DS_{ny} based on the y values, i.e., DS_i = Records in which $Y=v_i$
  - For each DS_i , learn a joint distribution of input distribution
  - This will give us p(X| $Y=v_i$), i.e., P(X_1, X_2, … X_m | $Y=v_i$)

Recipe for a Bayes Classifier

- Assume you want to predict output *Y* which has arity n_Y and values v_1, v_2, … v_{ny}.
- Assume there are *m* input attributes called X=(X_1, X_2, … X_m)
- Learn a conditional distribution of p(X|y) for each possible y value, y = v_1, v_2, … $v_{ny,}$, we do this by:
  - Break training set into n_Y subsets called DS_1, DS_2, … DS_{ny} based on the y values, i.e., DS_i = Records in which $Y=v_i$
  - For each DS_i , learn a joint distribution of input distribution
  - This will give us p(X| $Y=v_i$), i.e., P(X_1, X_2, … X_m | $Y=v_i$)

- Idea: When a new example ($X_1 = u_1$, $X_2 = u_2$, …. $X_m = u_m$) come along, predict the value of Y that has the highest value of P($Y=v_i$ | X_1, X_2, … X_m)

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

Getting what we need

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

Getting a posterior probability

$$P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$= \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$= \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v)}{\sum_{j=1}^{n_Y} P(X_1 = u_1 \cdots X_m = u_m \mid Y = v_j) P(Y = v_j)}$$

Bayes Classifiers in a nutshell

1. Learn the P(X_1, X_2, … X_m | $Y=v_i$) for each value v_i
3. Estimate P($Y=v_i$) as fraction of records with $Y=v_i$.
4. For a new prediction:

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$
$$= \underset{v}{\operatorname{argmax}}\, P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v)$$

Estimating the joint distribution of *X1*, *X2*, … X*m given y* can be problematic!

Joint Density Estimator Overfits

- Typically we don't have enough data to estimate the joint distribution accurately
- It is common to encounter the following situation:
  - If no records have the exact X=(u_1, u_2, …. u_m), then P(X|$Y=v_i$) = 0 for all values of Y.
- In that case, what can we do?
  - we might as well guess Y's value!

Example: Spam Filtering

- Bag-of-words representation is used for emails ($X = \{x_1, x_2, \ldots, x_m\}$)
- Assume that we have a dictionary containing all commonly used words and tokens
- We will create one attribute for each dictionary entry
  - E.g., x_i is a binary variable, x_i=1 (0) means the *i*th word in the dictionary is (not) present in the email
  - Other possible ways of forming the features exist, e.g., x_i=the # of times that the ith word appears
- Assume that our vocabulary contains10k commonly used words --- we have 10,000 attributes
- How many parameters that we need to learn?

$\boxed{2*(2^{10,000}-1)}$

- Clearly we don't have enough data to estimate that many parameters
- What can we do?
  - Make some bold assumptions to simplify the joint distribution

Naïve Bayes Assumption

- Assume that each attribute is independent of any other attributes given the class label

$$\boxed{\begin{aligned}&P(X_1 = u_1 \cdots X_m = u_m \mid Y = v_i)\\&= P(X_1 = u_1 \mid Y = v_i) \cdots P(X_m = u_m \mid Y = v_i)\end{aligned}}$$

A note about independence

- Assume A and B are Boolean Random Variables. Then

"A and B are independent"

if and only if

P(A|B) = P(A)

- "A and B are independent" is often notated as $A \perp B$

Independence Theorems

- Assume P(A|B) = P(A)
- Then P(A^B) =

= P(A) P(B)

- Assume P(A|B) = P(A)
- Then P(B|A) =

= P(B)

Independence Theorems

- Assume P(A|B) = P(A)
- Then P(~A|B) =

= P(~A)

- Assume P(A|B) = P(A)
- Then P(A|~B) =

= P(A)

Conditional Independence

- $P(X_1|X_2,y) = P(X_1|y)$
  - X_1 and X_2 are conditionally independent given y
- If X_1 and X_2 are conditionally independent given y, then we have
  - $P(X_1,X_2|y) = P(X_1|y) P(X_2|y)$

Naïve Bayes Classifier

- Assume you want to predict output *Y* which has arity n_Y and values v_1, v_2, … v_{ny}.
- Assume there are *m* input attributes called X=(X_1, X_2, … X_m)
- Learn a conditional distribution of p(X|y) for each possible y value, y = v_1, v_2, … v_{ny}, we do this by:
  - Break training set into n_Y subsets called DS_1, DS_2, … DS_{ny} based on the y values, i.e., DS_i = Records in which $Y=v_i$
  - For each DS_i , learn a joint distribution of input distribution

$$\boxed{\begin{aligned}&P(X_1 = u_1 \cdots X_m = u_m \mid Y = v_i)\\ &= P(X_1 = u_1 \mid Y = v_i) \cdots P(X_m = u_m \mid Y = v_i)\end{aligned}}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(X_1 = u_1 \mid Y = v) \cdots P(X_m = u_m \mid Y = v) P(Y = v)$$

Example

X_1	X_2	X_3	Y
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
0	0	1	1
0	1	1	1

Apply Naïve Bayes, and make prediction for (1,1,1)?

Final Notes about Bayes Classifier

- Any density estimator can be plugged in to estimate $P(X_1,X_2, \ldots, X_m | y)$
- Real valued attributes can be modeled using simple distributions such as Gaussian (Normal) distribution
- Zero probabilities are painful for both joint and naïve. A hack called Laplace smoothing can help!
- Naïve Bayes is wonderfully cheap and survives tens of thousands of attributes easily

What you should know

- Probability
  - Fundamentals of Probability and Bayes Rule
  - What's a Joint Distribution
  - How to do inference (i.e. P(E1|E2)) once you have a JD, using bayes rule
  - How to learn a Joint DE (nothing that simple counting cannot fix)
- Bayes Classifiers
  - What is a Bayes Classifier
  - What is a naïve bayes classifier, what is the naïve bayes assumption