Some references for probability

(We will be closest to chapter two of K&F—posted)

Wasserman, "All of statistics"
(Vision lab has a few copies that we can lend out).

Forsyth and Ponce chapter
http://luthuli.cs.uiuc.edu/~daf/book/bookpages/pdf/probability.pdf

Your favorite intro to probability book (e.g., "Mathematical statistics and Data Analysis," by John Rice.)

Google (and WikiPedia)

Probability review

Formulas that you should be very comfortable with are marked by *.

Interpretations of probability

1) Representation of expected frequency
2) Degree of belief

Basic terminology and rules

Space of outcomes (often denoted by Ω)

Event (subset of Ω)

Denote the space of measurable events (one we want to assign a probability to) by $\mathcal{S}$.
$\mathcal{S}$ must include $\varnothing$ and Ω
$\mathcal{S}$ is closed under set operations
$\alpha,\beta \in S \Rightarrow \alpha \cup \beta \in S$, $\alpha \cap \beta \in S$, $\alpha^C = \Omega - \alpha \in S$, etc.

Basic terminology and rules

A probability distribution P over $(\Omega,\mathcal{S})$ is a mapping from events in S to real values such that

If $a \in S, \quad P(a) \geq 0$

$P(\Omega) = 1$

If $\alpha, \beta \in S$, and $a \cap b = \varnothing$, then $P(a \cup b) = P(a) + P(b)$

Basic terminology and rules

We can further derive additional familiar facts

If $a \in S, \quad P(a) \in [0,1]$

$P(\varnothing) = 0$

If $\alpha, \beta \in S$, then $P(a \cup b) = P(a) + P(b) - P(a \cap b)$

The probabilities over disjoint sets that cover P(Ω) sum to 1.

Basic terminology and rules

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Example, what is the probability that you have rolled 2, given that you know you have rolled a prime number?

Basic terminology and rules

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a bit of algebra,

$$P(A \cap B) = P(A)P(B|A)$$

In general, we have the chain (product) rule

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$ *

$$P(A_1 \cap A_2 \cap \ \dots \ A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ \dots \ P(A_N|A_1 \cap A_2 \cap \ \dots \ A_{N-1})$$

Basic terminology and rules

From before, we define conditional probability by

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

$$P(A \cap B) = P(A)P(B|A)$$

and $P(A \cap B) = P(B)P(A|B)$

and thus $P(B)P(A|B) = P(A)P(B|A)$

and we get $P(A|B) = \dfrac{P(A)P(B|A)}{P(B)}$ Bayes rule *

Example

Probability of disease given symptoms

Suppose a TB test is 95% accurate

$P(\text{positive} \mid TB) = 0.95$

$P(\text{negative} \mid \widetilde{TB}) = 0.95$

What is P (TB | positive)?

Example (continued)

Probability of disease given symptoms

Suppose a TB test is 95% accurate
Suppose that TB is in 0.1% of population

What is P (TB | positive)?

Example (continued)

$P(TB \mid positive)$

$$= \frac{P(positive \mid TB)P(TB)}{P(positive)}$$

$$= \frac{P(positive \mid TB)P(TB)}{P(positive \mid TB)P(TB) + P(positive \mid \widetilde{TB})P(\widetilde{TB})}$$

$P(TB \mid positive)$

$$= \frac{P(positive \mid TB)P(TB)}{P(positive)}$$

$$= \frac{P(positive \mid TB)P(TB)}{P(positive \mid TB)P(TB) + P(positive \mid \widetilde{TB})P(\widetilde{TB})}$$

$$= \frac{(0.95)(0.001)}{(0.95)(0.001) + (0.05)(0.999)}$$

$$\cong 0.0187$$

Basic rules (so far)

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Chain (Product) Rule

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$ *

$$P(A_1 \cap A_2 \cap \;\dots\; A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \;\dots\; P(A_N|A_1 \cap A_2 \cap \;\dots\; A_{N-1})$$

Bayes Rule

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

Random Variables

Random variables

- Defined by functions mapping outcomes to values
- By choice, whatever we are interested in
- Typically denoted by uppercase letters (e.g., X)
- Generic values are corresponding lower case letters
- Shorthand: P(x) = P(X=x)
- Value "type" is arbitrary (typically categorical or real)

Example (from K&F)

- Outcomes are student grades (A,B,C)
- Random variable G=f_{GRADE}(student)
- $P(A) \equiv P(G = A) \equiv P(\{w \in \Omega : f_{GRADE}(w) = A\})$

Joint Distributions of Random Variables

Joint distribution of random variables

$$P(X,Y) \equiv P(X = x, Y = y) \equiv P(\{w \in \Omega : X(w) = x \text{ and } Y(w) = y\})$$

Conditional definition, Bayes rule, chain rule all apply.

Marginal distributions ("sum rule")

$$P(X) = \sum_Y P(X,Y)$$ *

Chain (product) rule (two variable case of chain rule)

$$P(X,Y) = P(X\,|\,Y)P(Y)$$ *

Basic terminology and rules

Conditional probability

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

Bayes

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

$P(X\,|\,Y) \propto P(Y|X)P(X)$ (when Y is constant, i.e., evidence) *

Normalization

Often we will deal with quantities or functions which are proportional to probabilities (OK if we just want argmax)

To convert such quantities to probabilities we *normalize*.

$$\text{if} \quad p(x) \propto P(X=x) \quad \text{then} \quad P(X=x) = \frac{p(x)}{\sum_x p(x)}$$

Example: $P(X \mid Y) \propto P(X,Y)$

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

Independence

This can cause confusion. If P(Y) is zero, the other case cannot be used (divide by zero). However, in this case, Y never happens, and so we (a priori) have a choice to declare whether X is independent from Y or not. However, under scrutiny, the choice does make sense, and allows consistency with the second definition. Note that the second formula works in this (weird) case because if P(Y)=0, then P(X,Y) is also 0.

$$X \perp Y \quad \Leftrightarrow \quad P(X \mid Y) = P(X) \quad \text{or} \quad \text{P(Y)=0}$$ *

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y)$$ *

Note that Bishop uses $\perp\!\!\!\perp$ instead of $\perp$

Conditional Independence

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X \mid Y,Z) = P(X \mid Z) \quad \text{or} \quad \text{P(Y,Z)=0}$$ *

Equivalent, sometimes more convenient definition

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$ *