10-601: Recitation 1
Math Review and Decision Trees

January 15, 2015

1 Math background

1.1 Probability and statistics

1.1.1 Axioms of probability

1. $P(A) \geq 0$, where A is an event.
2. $P(\Omega) = 1$, where Ω is the sample space.
3. $P(A \cup B) = P(A) + P(B)$, for *disjoint* A, B.

1.1.2 Set operations apply to probabilities

1. $$\begin{aligned} \textbf{Commutative}: & A \cup B = B \cup A \\ & A \cap B = B \cap A \end{aligned}$$
2. $$\begin{aligned} \textbf{Associative}: & A \cup (B \cup C) = (A \cup B) \cup C \\ & A \cap (B \cap C) = (A \cap B) \cap C \end{aligned}$$
3. $$\begin{aligned} \textbf{Distributive}: & A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \\ & A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \end{aligned}$$
4. $$\begin{aligned} \textbf{DeMorgan's}: & (A \cup B)^c = A^c \cap B^c \\ & (A \cap B)^c = A^c \cup B^c \end{aligned}$$

1.1.3 Conditional probabilities and independence

The *conditional probability of A given B* is $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (assuming $P(B) > 0$).
We can rearrange this definition to get the chain rule for probabilities: $P(A \cap B) = P(A|B)P(B)$.

Example: We roll two fair dice. What is the probability that the sum is greater than six, given that one die is a three?

A, B are *independent* if $P(A \cap B) = P(A)P(B)$; equivalently, if $P(B|A) = P(B)$ (when $P(A) > 0$).

A, B are *conditionally independent given* C if $P(A \cap B|C) = P(A|C)P(B|C)$.

1.1.4 Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{1}$$

Example: Suppose you're being screened for a rare, but serious disease that only 5% of the population has. The screening procedure is 99% accurate, meaning it returns a positive result for 99% of people who have the disease. If a person does not have the disease, the test returns positive 0.1% of the time. Your test returned a positive result. What is the chance that you actually have this disease; that is, what is $P(D = true|T = positive)$?

1.1.5 Some definitions

Expectation/mean:

$$\mu = E(X) = \begin{cases} \sum_x xp(x), & X \text{ discrete} \\ \int_{-\infty}^{\infty} xp(x)dx, & X \text{ continuous} \end{cases} \tag{2}$$

Variance:

$$\sigma^2 = Var(X) = E((X - \mu)^2) = E(X^2) - \mu^2 \tag{3}$$

For a sample $X_1, ..., X_n$, the sample mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$ and sample variance $S^2 = \frac{1}{n} \sum_i (X_i - \bar{X}_n)^2$.

Entropy:

$$H(Y) = -\sum_y P(Y = y) \log_2 P(Y = y) \tag{4}$$

Conditional entropy:

$$\begin{aligned} H(Y|X_i) &= \sum_x P(X_i = x) H(Y|X_i = x) \\ &= -\sum_x P(X_i = x) \sum_y P(Y = y|X_i = x) \log_2 P(Y = y|X_i = x) \end{aligned} \tag{5}$$

Mutual information / information gain:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{6}$$

1.1.6 Law of large numbers and central limit theorem

Law of large numbers: If $X_1, ..., X_n$ are iid with mean μ, then $\bar{X}_n \to \mu$, as $n \to \infty$.

Central limit theorem: If $X_1, ..., X_n$ are iid with mean μ and variance σ^2, then $\sqrt{n}(\bar{X}_n - \mu) \to \mathcal{N}(0, \sigma^2)$, as $n \to \infty$. (There are lots of other ways to express the CLT; *e.g.*, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathcal{N}(0, 1)$, $\bar{X}_n \to \mathcal{N}(\mu, \sigma^2/n)$, etc.)

1.2 Linear algebra

Be familiar with: matrix/vector operations, determinants, norms, linear independence, ...

(lots of good resources listed on homework 1)

1.3 Calculus

Make sure you can... differentiate functions wrt variables (important for optimization).

(lots of good resources listed on homework 1)

2 Decision trees

Let's run the ID3 algorithm on a short example to build a decision tree to predict the habitability of planets based on size and orbit (question courtesy of Martin Azizyan).

Table 1: Planet size, orbit, and habitability

Size	Orbit	Habitable	Count
big	near	yes	20
big	far	yes	170
small	near	yes	139
small	far	yes	45
big	near	no	130
big	far	no	30
small	near	no	11
small	far	no	255