Machine Learning Cheatsheet

Janet Matsen's Machine Learning (ML) notes from CSE 446, Winter 2016. `http://courses.cs.washington.edu/courses/cse446/16wi/`
Used LaTeX template from an existing Statistics cheat sheet: `https://github.com/wzchen/probability_cheatsheet`, by William Chen (`http://wzchen.com`) and Joe Blitzstein.

Last Updated January 10, 2016

Math/Stat Review

Random Variable X belongs to set Ω

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability. $P(A|B) = P(A \cap B)/P(B)$

Bayes' Rule - Bayes' Rule unites marginal, joint, and conditional probabilities. We use this as the definition of conditional probability.

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

$$P(A = a \mid B) = \frac{P(A = a)P(B \mid A = a)}{\sum_{a'} P(A = a)P(B \mid A = a)}$$

Law of Total Probability : $\sum_x P(X = x) = 1$

Product Rule : $P(A, B) = P(A \mid B) \cdot P(B)$

Sum Rule : $P(A) = \sum_{x \in \Omega} P(A, B = b)$

Law of Total Probability (LOTP)

Let $B_1, B_2, B_3, ...B_n$ be a *partition* of the sample space (i.e., they are disjoint and their union is the entire sample space).

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

For **LOTP with extra conditioning**, just add in another event C!

$$P(A|C) = P(A|B_1, C)P(B_1|C) + \cdots + P(A|B_n, C)P(B_n|C)$$
$$P(A|C) = P(A \cap B_1|C) + P(A \cap B_2|C) + \cdots + P(A \cap B_n|C)$$

Special case of LOTP with B and B^c as partition:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Bayes' Rule

Bayes' Rule, and with extra conditioning (just add in C!)

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

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

We can also write

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

Odds Form of Bayes' Rule

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

The *posterior odds* of A are the *likelihood ratio* times the *prior odds*.

Practice: What is $P(disease \mid +test)$ if P(disease) = 0.01, P(+ | disease) = 0.99, P(+ | no disease) = 0.01?

Expectation

f(X) probability distribution function of X

X ∼ P : X is distributed according to P.

Expected value of f under P : $E_P[f(x)] = \sum_x p(x)f(x)$

E.g. unbiased coin. x = 1, 2, 3, 4, 5, 6. p(X=x) = 1/6 for all x.
E(X) = $\sum_x p(x) \cdot x = (1/6) \cdot [1 + 2 + 3 + 4 + 5 + 6] = 3.5$

Entropy

$X \sim P$, $x \in \Omega$

First define **Surprise**: $S(x) = -\log_2 p(x)$
$S(X = \text{heads}) = -\log_2(1/2) = 1$.

Axiom 1 : S(1) = 0. (If an event with probability 1 occurs, it is not surprising at all.)

Axiom 2 : S(q) > S(p) if q < p. (When more unlikely outcomes occur, it is more surprising.)

Axiom 3 : S(p) is a continuous function of p. (If an outcomes probability changes by a tiny amount, the corresponding surprise should not change by a big amount.)

Axiom 4 : S(pq) = S(p) + S(q). (Surprise is additive for independent outcomes.)

Surprise of 7 = pretty surprised. Probability of $1/2^7$ of happening (Shannon) **Entropy**:

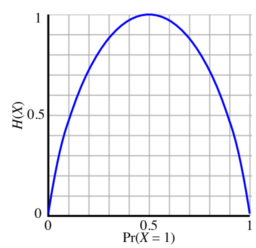
$$\begin{aligned} H[X] &= -\sum_x p(x) \cdot \log_2 p(x) \\ &= -\sum_x p(x)S(x) \\ &= E[S(x)] \end{aligned}$$

The entropy is the expectation of the surprise. Throw out x for $p(x) = 0$ because log(0) is ∞.

<u>Entropy of an unbiased coin flip:</u>
X is a coin flip. $P(X = \text{heads}) = 1/2$, $P(X = \text{tails}) = 1/2$
Note: $\log_2(1/2) = -1$, $-\log_2(1/2) = \log_2(2) = 1$
$H[X] = -[1/2 \log_2(1/2) + 1/2 \log_2(1/2)] = 1$

<u>Entropy of a coin that always flips to heads:</u>
$P(X = \text{heads}) = 1$, $P(X = \text{tails}) = 0$
Note: $\log_x(0) = 0$
$H[X] = -[1 \log_2(1) + 0] = 0$
No surprise: you are sure what you are going to get.

Binary entropy plot.



<u>Canonical example:</u>

X	Y
0	1
1	0
1	1

If you want to estimate entropy of X, you can use P(X=0).

$$\begin{aligned} H[X] &= -[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}] \\ &= \frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 3 - \frac{2}{3}\log_2 2 \\ &= \log_2 3 - \frac{2}{3} \approx 0.91 \end{aligned}$$

This time H[X] = H[Y] because of symmetry.

Conditional Entropy

If you don't know x: (this is kind of an average).
$H[Y \mid X = x] = -\sum_y P(Y = y \mid X = x) \cdot \log_2 P(y \mid X = x)$
$H[Y \mid X = x] = E[S(Y \mid X = x)]$
Note that we are summing over y because we are specifying x.

For a particular value of X:
$H[Y \mid X] = \sum_x p(x)H[Y \mid X = x]$

Back to table above:

$$H[Y \mid X = 0] = ?$$
look only at X=0 in table.
$$= -[0 + 1 \log_2]$$

Now that you know X=0, entropy goes to 0.

$H[Y \mid X = 1] = 1$: You know *less* if you know X=1.

Now use $H[Y \mid X] = \frac{1}{3}(0) + \frac{2}{3}(1) = 2/3$
Given X, you know more. Average our the more certain case and the less certain case.

Note: $H[Y \mid X] \leq H[Y]$: knowing something can't make you know less.

Decison Trees

Vocab

- decision tree

Let's do this thing.