Probabilty Notes

Yychi Fyu

@SIST, ShanghaiTech University

December 19, 2019

Odds: the *odds* of an event A are

$$\text{odds}(A) = P(A)/P(A^c).$$

For example, if $P(A) = 2/3$, we say the odds in favor of A are 2 to 1. (This is sometimes written as $2 : 1$, and is sometimes stated as 1 to 2 odds against A; care is needed since some sources do not explicitly state whether they are referring to odds in favor or odds against an event). Of course we can also convert from odds back to probability:

$$P(A) = \text{odds}(A)/(1 + \text{odds}(A)).$$

By the way, the log of odds is known as logits, denote $P(A)$ as p, the logits of p is

$$logits(p) = \log \frac{p}{1-p},$$

which is the inverse of sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}} = logits^{-1}.$$

You must have met these in logistic regression. Both odds and logits are increasing as p increases.

(Odds form of Bayes' rule): For any events A and B with positive probabilities, the odds of A after conditioning on B are

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

In words, this says that the *posterior odds* $P(A|B)/P(A^c|B)$ are equal to the *prior odds* $P(A)/P(A^c)$ times the factor $P(B|A)/P(B|A^c)$, which is known in statistics as the *likelihood ratio*.

Superise: The *surprise* of learning that an event with probability p happened is defined as $\log_2(1/p)$, measured in a unit called bits. Low-probability events have high surprise, while an event with probability 1 has zero surprise. The log is there so that if we observe two independent events A and B, the total surprise is the same as the surprise from observing $A \cap B$. The log is base 2 so that if we learn that an event with probability $1/2$ happened, the surprise is 1, which corresponds to having received 1 bit of information.

Entropy: Let X be a discrete r.v. whose distinct possible values are $a_1, a_2, \ldots, a_n$, with probabilities $p_1, p_2, \ldots, p_n$ respectively (so $p_1 + p_2 + \cdots + p_n = 1$). The *entropy* of X is defined to be the average surprise of learning the value of X:

$$H(X) = \sum_{j=1}^{n} p_j \log_2(1/p_j).$$

Kullback-Leibler divergence: Let $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{r} = (r_1, \ldots, r_n)$ be two probability vectors (so each is nonnegative and sums to 1). Think of each as a possible PMF for a random variable whose support consists of n distinct values. The *Kullback-Leibler divergence* between $\mathbf{p}$ and $\mathbf{r}$ is defined as

$$D(\mathbf{p}, \mathbf{r}) = \sum_{j=1}^{n} p_j \log_2(1/r_j) - \sum_{j=1}^{n} p_j \log_2(1/p_j).$$

This is the difference between the average surprise we will experience when the actual probabilities are **p** but we are instead working with **r** (for example, if **p** is unknown and **r** is our current guess for **p**), and our average surprise when we work with **p**.