Probabilty Notes

Yychi Fyu
@SIST, ShanghaiTech University

July 20, 2020

Definition (Odds). The *odds* of an event A are

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

For[1] 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.

Definition (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*.

Definition (Surprise). 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.

Definition (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:

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

Definition (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**.

1 Probabilty

1.1 Bounds on tail probability

Theorem 1.1 (Markov's inequality). *For any r.v.* X *and constant* $a > 0$,

$$P(|X| \geq a) \leq \frac{E|X|}{a}. \tag{1.1}$$

Proof. Let $Y = \frac{|X|}{a}$. We need to show that $P(Y \geq 1) \leq E(Y)$. Note that

$$I(Y \geq 1) \leq Y,$$

since if $I(Y \geq 1) = 0$ then $Y \geq 0$, and if $I(Y \geq 1) = 1$ then $Y \geq 1$ (because the indicator says so). Taking the expectation of both sides, we have Markov's inequality. □

Theorem 1.2 (Chebyshev's inequality). *Let* X *have mean* μ *and variance* σ^2*. Then for any* $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}. \tag{1.2}$$

Proof of Chebyshev's inequality. By Markov's inequality (Theorem 1.1),

$$P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2) \leq \frac{E(X - \mu)^2}{a^2} = \frac{\sigma^2}{a^2}.$$

□

Theorem 1.3 (Chernoff bound). *For any r.v.* X *and constants* $a > 0$ *and* $t > 0$,

$$P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}. \tag{1.3}$$

Proof. The transformation g with $g(x) = e^{tx}$ is invertible and strictly increasing. So by Markov's inequality, we have

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}.$$

□

1.2 Law of large numbers

Assume we have i.i.d. $X_1, X_2, X_3, \dots$ with finite mean μ and finite variance σ^2. For all positive integers n, let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

be the *sample mean* of X_1 through X_n. The sample mean is itself an r.v., with mean μ and variance σ^2/n:

$$E(\bar{X}_n) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu,$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n}.$$

Theorem 1.4 (Strong law of large numbers). *The sample mean $\bar{X}_n$ converges to the true mean μ pointwise as $n \to \infty$, with probability 1. In other words,*

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1, \text{ or } \bar{X}_n \xrightarrow{a.s.} \mu. \tag{1.4}$$

Theorem 1.5 (Weak law of large numbers). *For all $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$. (This is called* convergence in probability.*) In other words,*

$$\lim_{n\to\infty} P(\bar{X}_n = \mu) = 1. \tag{1.5}$$

Proof. Fix $\epsilon > 0$, by Chebyshev's inequality,

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

As $n \to \infty$, the right-hand side goes to 0, ans so must the left-hand side. □

2 Information theory

hi here, /usr/bin/locale, seems good haha, are you kidding me? yes i can

References

[1] J. K. Blitzstein and J. Hwang, *Introduction to probability.* Crc Press, 2019.