CHAPTER

PROBABILITY INEQUALITIES

8

CHAPTER CONTENTS

If probability mass/density function $f(x)$ is given explicitly, the values of the probability (density) can be computed. However, in reality, $f(x)$ itself may not be given explicitly, but only partial information such as the expectation $E[x]$ or the variance $V[x]$ is given. In this chapter, various inequalities are introduced that can be used for evaluating the probability only from partial information. See [19] for more details.

8.1 UNION BOUND

Let us recall the additive law of probabilities shown in Section 2.2:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Since $\Pr(A \cap B)$ is non-negative, the following inequality is immediately obtained:

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B),$$

which is called the *union bound*. Even if $\Pr(A \cup B)$ is difficult to obtain explicitly, the union bound gives its upper bound from the probabilities of each event. The union

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00019-4

bound can be extended to multiple events: for $A_1, \ldots, A_N$,

$$\Pr(A_1 \cup \cdots \cup A_N) \leq \Pr(A_1) + \cdots + \Pr(A_N).$$

8.2 INEQUALITIES FOR PROBABILITIES

In this section, inequalities for probabilities based on the expectation and variance are introduced.

8.2.1 MARKOV'S INEQUALITY AND CHERNOFF'S INEQUALITY

For *non-negative* random variable x having expectation $E[x]$,

$$\Pr\big(x \geq a\big) \leq \frac{E[x]}{a} \tag{8.1}$$

holds for any positive scalar a (Fig. 8.1). This is called *Markov's inequality*, which allows us to know the upper bound of the probability only from the expectation. Since $\Pr(x < a) = 1 - \Pr(x \geq a)$, a lower bound can also be obtained similarly:

$$\Pr\big(x < a\big) \geq 1 - \frac{E[x]}{a}.$$

Markov's inequality can be proved by the fact that the function

$$g(x) = \begin{cases} a & (x \geq a), \\ 0 & (0 \leq x < a), \end{cases}$$

defined for $x \geq 0$ satisfies $x \geq g(x)$:

$$E[x] \geq E[g(x)] = a \Pr(x \geq a).$$

For arbitrary non-negative and monotone increasing function $\phi(x)$, Markov's inequality can be generalized as

$$\Pr(x \geq a) = \Pr\big(\phi(x) \geq \phi(a)\big) \leq \frac{E[\phi(x)]}{\phi(a)}. \tag{8.2}$$

Setting $\phi(x) = e^{tx}$ for $t > 0$ in Eq. (8.2) yields

$$\Pr(x \geq a) = \Pr\big(e^{tx} \geq e^{ta}\big) \leq \frac{E[e^{tx}]}{e^{ta}}, \tag{8.3}$$

which is called *Chernoff's inequality*. Minimizing the right-hand side of (8.3) with respect to t yields a tighter upper bound.

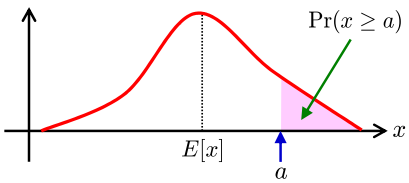

FIGURE 8.1

Markov's inequality.

8.2.2 CANTELLI'S INEQUALITY AND CHEBYSHEV'S INEQUALITY

Markov's inequality upper-bounds the probability based on the expectation $E[x]$. Here, upper bounds of the probability based on the variance $V[x]$ in addition to the expectation $E[x]$ are introduced.

When a random variable x possesses the expectation $E[x]$ and variance $V[x]$, the generic inequality (coming from $a \geq b \Longrightarrow a^2 \geq b^2$)

$$\Pr(a \geq b) \leq \Pr(a^2 \geq b^2)$$

and Markov's inequality (8.1) yield that for a positive scalar ε,

$$\begin{aligned}\Pr\Big(x - E[x] \geq \varepsilon\Big) &= \Pr\Big(\varepsilon(x - E[x]) + V[x] \geq V[x] + \varepsilon^2\Big) \\ &\leq \Pr\Big(\{\varepsilon(x - E[x]) + V[x]\}^2 \geq \{V[x] + \varepsilon^2\}^2\Big) \\ &\leq \frac{E\big[\{\varepsilon(x - E[x]) + V[x]\}^2\big]}{\{V[x] + \varepsilon^2\}^2} = \frac{V[x]}{V[x] + \varepsilon^2}.\end{aligned}$$

This is called *Cantelli's inequality* or *one-sided Chebyshev's inequality*. Similarly, the following inequality also holds:

$$\Pr\Big(x - E[x] \leq -\varepsilon\Big) \leq \frac{V[x]}{V[x] + \varepsilon^2}.$$

Furthermore, Markov's inequality (8.1) yields

$$\Pr\Big(|x - E[x]| \geq \varepsilon\Big) = \Pr\Big((x - E[x])^2 \geq \varepsilon^2\Big) \leq \frac{V[x]}{\varepsilon^2}, \tag{8.4}$$

which is called *Chebyshev's inequality* (Fig. 8.2). While Markov's inequality can only bound one-sided probabilities, Chebyshev's inequality allows us to bound two-sided probabilities. A lower bound can also be obtained similarly:

$$\Pr\Big(|x - E[x]| < \varepsilon\Big) \geq 1 - \frac{V[x]}{\varepsilon^2}. \tag{8.5}$$

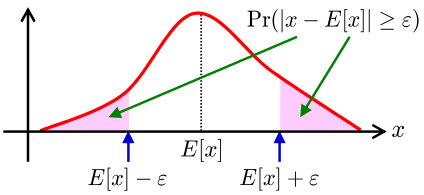

FIGURE 8.2

Chebyshev's inequality.

Chebyshev's inequality can be extended to an arbitrary interval $[a, b]$ as

$$\Pr\Big(a < x < b\Big) \geq 1 - \frac{V[x] + \Big(E[x] - \frac{a+b}{2}\Big)^2}{\Big(\frac{b-a}{2}\Big)^2},$$

which can be proved by applying Markov's inequality as

$$\begin{aligned}\Pr\Big((x \leq a) \cup (b \leq x)\Big) &= \Pr\left(\left|x - \frac{a+b}{2}\right| \geq \frac{b-a}{2}\right) \\ &= \Pr\left(\left(x - \frac{a+b}{2}\right)^2 \geq \left(\frac{b-a}{2}\right)^2\right) \\ &\leq \frac{E[(x - \frac{a+b}{2})^2]}{\Big(\frac{b-a}{2}\Big)^2} = \frac{V[x] + \Big(E[x] - \frac{a+b}{2}\Big)^2}{\Big(\frac{b-a}{2}\Big)^2}.\end{aligned}$$

Note that the above inequality is reduced to original Chebyshev's inequality (8.5) by setting

$$a = -\varepsilon + E[x] \quad \text{and} \quad b = \varepsilon + E[x].$$

8.3 INEQUALITIES FOR EXPECTATION

In this section, inequalities for the expectation are introduced.

8.3.1 JENSEN'S INEQUALITY

For all $\theta \in [0, 1]$ and all $a < b$, a real-valued function $h(x)$ that satisfies

$$h\big(\theta a + (1-\theta)b\big) \leq \theta h(a) + (1-\theta)h(b) \tag{8.6}$$

is called a *convex function* (see Fig. 8.3).

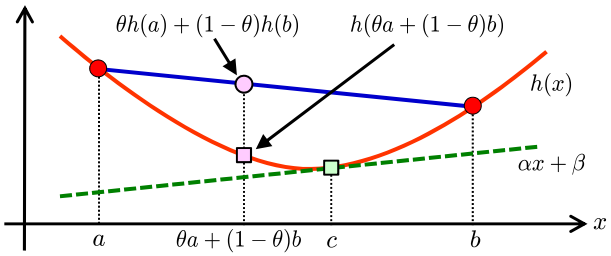

FIGURE 8.3

Convex function and tangent line.

If $h(x)$ is convex, for any c, there exists a tangent line

$$g(x) = \alpha x + \beta$$

such that it touches $h(x)$ at c and lower-bounds $h(x)$ (Fig. 8.3):

$$g(c) = h(c) \quad \text{and} \quad g(x) \le h(x) \quad \text{for all } x.$$

Setting $c = E[x]$ yields

$$E\big[h(x)\big] \ge E\big[g(x)\big] = \alpha E[x] + \beta = g(E[x]) = h(E[x]),$$

which is called *Jensen's inequality*. In practice, computing the expectation $E\big[h(x)\big]$ is often hard due to nonlinear transformation $h(x)$. On the other hand, computing $h(E[x])$ may be straightforward because it is just a nonlinear transformation of the expectation. Thus, Jensen's inequality allows us to know a lower bound of $E\big[h(x)\big]$ even if it is hard to compute.

Jensen's inequality can be extended to multidimensional convex function $h(\boldsymbol{x})$:

$$E\big[h(\boldsymbol{x})\big] \ge h(E[\boldsymbol{x}]).$$

8.3.2 HÖLDER'S INEQUALITY AND SCHWARZ'S INEQUALITY

For scalars p and q such that

$$\frac{1}{p} + \frac{1}{q} = 1,$$

if random variables $|x|^p$ and $|y|^q$ possess the expectations, the following *Hölder's inequality* holds:

$$E\big[|xy|\big] \le \left(E\big[|x|^p\big]\right)^{1/p} \left(E\big[|y|^q\big]\right)^{1/q}. \tag{8.7}$$

Hölder's inequality can be proved as follows. For $0 \le \theta \le 1$, setting $h(x) = e^x$ in Eq. (8.6) yields

$$e^{\theta a+(1-\theta)b} \le \theta e^a + (1-\theta)e^b.$$

Setting

$$\theta = \frac{1}{p}, \quad 1-\theta = \frac{1}{q}, \quad a = \log \frac{|x|^p}{E\left[|x|^p\right]}, \quad \text{and} \quad b = \log \frac{|y|^q}{E\left[|y|^q\right]}$$

yields

$$\frac{|xy|}{\left(E\left[|x|^p\right]\right)^{1/p}\left(E\left[|y|^q\right]\right)^{1/q}} \le \frac{1}{p}\frac{|x|^p}{E\left[|x|^p\right]} + \frac{1}{q}\frac{|y|^q}{E\left[|y|^q\right]}.$$

Taking the expectation of both sides yields Eq. (8.7) because the expectation of the right-hand side is 1.

Hölder's inequality for $p = q = 2$ is particularly referred to as *Schwarz's inequality*:

$$E\left[|xy|\right] \le \sqrt{E\left[|x|^2\right]}\sqrt{E\left[|y|^2\right]}.$$

8.3.3 MINKOWSKI'S INEQUALITY

For $p \ge 1$, *Minkowski's inequality* is given by

$$\left(E\left[|x+y|^p\right]\right)^{1/p} \le \left(E\left[|x|^p\right]\right)^{1/p} + \left(E\left[|y|^p\right]\right)^{1/p}. \tag{8.8}$$

Minkowski's inequality can be proved as follows. A generic inequality

$$|x+y| \le |x| + |y|$$

yields

$$E\left[|x+y|^p\right] \le E\left[|x| \cdot |x+y|^{p-1}\right] + E\left[|y| \cdot |x+y|^{p-1}\right].$$

When $p = 1$, this immediately yields Eq.(8.8). When $p > 1$, applying Hölder's inequality to each term in the right-hand side yields

$$E\left[|x| \cdot |x+y|^{p-1}\right] \le \left(E\left[|x|^p\right]\right)^{1/p}\left(E\left[|x+y|^{(p-1)q}\right]\right)^{1/q},$$
$$E\left[|y| \cdot |x+y|^{p-1}\right] \le \left(E\left[|y|^p\right]\right)^{1/p}\left(E\left[|x+y|^{(p-1)q}\right]\right)^{1/q},$$

where $q = \frac{p}{p-1}$. Then

$$E\left[|x+y|^p\right] \le \left(\left(E\left[|x|^p\right]\right)^{1/p} + \left(E\left[|y|^p\right]\right)^{1/p}\right)\left(E\left[|x+y|^p\right]\right)^{1-1/p}$$

holds, and dividing the both side by $\left(E\left[|x+y|^p\right]\right)^{1-1/p}$ yields Eq. (8.8).

8.3.4 KANTOROVICH'S INEQUALITY

For random variable x such that $0 < a \leq x \leq b$, *Kantorovich's inequality* is given by

$$E[x]E\left[\frac{1}{x}\right] \leq \frac{(a+b)^2}{4ab}. \tag{8.9}$$

Kantorovich's inequality can be proved as follows. A generic inequality

$$0 \leq (b-x)(x-a) = (a+b-x)x - ab$$

yields

$$\frac{1}{x} \leq \frac{a+b-x}{ab},$$

which yields

$$E[x]E\left[\frac{1}{x}\right] \leq \frac{E[x](a+b-E[x])}{ab}.$$

Completing the square as

$$\begin{aligned} E[x](a+b-E[x]) &= -\left(E[x] - \frac{a+b}{2}\right)^2 + \frac{(a+b)^2}{4} \\ &\leq \frac{(a+b)^2}{4} \end{aligned}$$

yields

$$\begin{aligned} E[x]E\left[\frac{1}{x}\right] &\leq \frac{-(E[x]-(a+b)/2)^2 + (a+b)^2/4}{ab} \\ &\leq \frac{(a+b)^2}{4ab}, \end{aligned}$$

which proves Eq. (8.9).

8.4 INEQUALITIES FOR THE SUM OF INDEPENDENT RANDOM VARIABLES

In this section, inequalities for the sum and average of independent random variables $x_1, \ldots, x_n$,

$$\widetilde{x} = \sum_{i=1}^{n} x_i \quad \text{and} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

are introduced.

8.4.1 CHEBYSHEV'S INEQUALITY AND CHERNOFF'S INEQUALITY

For $\widetilde{x} - E[\widetilde{x}]$, Chebyshev's inequality (8.4) yields

$$\begin{aligned}\Pr\left(|\widetilde{x} - E[\widetilde{x}]| \geq \varepsilon\right) &= \Pr\left((\widetilde{x} - E[\widetilde{x}])^2 \geq \varepsilon^2\right) \\ &\leq \frac{V[\widetilde{x}]}{\varepsilon^2} = \frac{\sum_{i=1}^{n} V[x_i]}{\varepsilon^2}.\end{aligned}$$

When $V[x_1] = \cdots = V[x_n] = \sigma^2$, this yields

$$\Pr\left(|\overline{x} - E[\overline{x}]| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

This upper bound is proportional to $1/n$.

Similarly, for arbitrary positive t, Chernoff's inequality (8.3) yields

$$\begin{aligned}\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) &\leq \exp\left(-t\varepsilon\right) E\left[\exp\left(t\sum_{i=1}^{n}(x_i - E[x_i])\right)\right] \\ &= \exp\left(-t\varepsilon\right)\prod_{i=1}^{n} E\left[\exp\left(t(x_i - E[x_i])\right)\right]. \qquad (8.10)\end{aligned}$$

This upper bound is the product of the moment-generating functions of $x_i - E[x_i]$ for $i = 1, \ldots, n$, and therefore it is expected to decrease exponentially with respect to n.

8.4.2 HOEFFDING'S INEQUALITY AND BERNSTEIN'S INEQUALITY

For random variables x_i such that $a_i \leq x_i \leq b_i$ for $i = 1, \ldots, n$, applying Hoeffding's formula,

$$E\left[\exp\left(t(x_i - E[x_i])\right)\right] \leq \exp\left(\frac{t^2(b_i - a_i)^2}{8}\right),$$

to Chernoff's inequality (8.10) yields

$$\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) \leq \exp\left(\frac{t^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - t\varepsilon\right).$$

Setting

$$t = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

to minimize the above upper bound yields

$$\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

This is called *Hoeffding's inequality*. Its variant for sample average $\overline{x}$ is given as

$$\Pr\left(\overline{x} - E[\overline{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

For random variables x_i such that $|x_i - E[x_i]| \leq a$ for $i = 1, \ldots, n$, *Bernstein's inequality* is given as

$$\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^{n} V[x_i] + 2a\varepsilon/3}\right).$$

Its derivation will be explained in Section 8.4.3. Its variant for sample average $\overline{x}$ is given as

$$\Pr\left(\overline{x} - E[\overline{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{\frac{2}{n}\sum_{i=1}^{n} V[x_i] + 2a\varepsilon/3}\right).$$

When $V[x_1] = \cdots = V[x_n] = \varepsilon$, this yields

$$\Pr\left(\overline{x} - E[\overline{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon}{2 + 2a/3}\right).$$

Thus, for small positive ε, Bernstein's inequality $\exp(-n\varepsilon)$ gives a tighter upper bound than Hoeffding's inequality $\exp(-n\varepsilon^2)$. This is because Bernstein's inequality uses the variance of $V[x_i]$, while Hoeffding's inequality only uses the domain $[a_i, b_i]$ of each random variable x_i.

8.4.3 BENNETT'S INEQUALITY

For random variables x_i such that $|x_i - E[x_i]| \leq a$ for $i = 1, \ldots, n$, applying Bennett's formula,

$$E\left[\exp\left(t(x_i - E[x_i])\right)\right] \leq \exp\left(V[x_i]\frac{\exp(ta) - 1 - ta}{a^2}\right),$$

to Chernoff's inequality (8.10) yields

$$\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) \leq \exp\left(\sum_{i=1}^{n} V[x_i]\frac{\exp(ta) - 1 - ta}{a^2} - t\varepsilon\right).$$

Setting

$$t = \frac{1}{a}\log\left(\frac{a\varepsilon}{\sum_{i=1}^{n} V[x_i]} + 1\right)$$

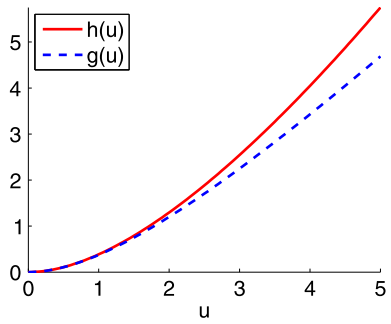

FIGURE 8.4

$h(u) = (1+u)\log(1+u) - u$ and $g(u) = \frac{u^2}{2+2u/3}$.

to minimize the above upper bound yields

$$\Pr\left(\widetilde{x} - E[\widetilde{x}] \geq \varepsilon\right) \leq \exp\left(-\frac{\sum_{i=1}^{n} V[x_i]}{a^2} h\left(\frac{a\varepsilon}{\sum_{i=1}^{n} V[x_i]}\right)\right),$$

where

$$h(u) = (1+u)\log(1+u) - u.$$

This is called *Bennett's inequality*.

For $u \geq 0$, the following inequality holds (Fig. 8.4):

$$h(u) \geq g(u) = \frac{u^2}{2 + 2u/3}.$$

Further upper-bounding Bennett's inequality by this actually gives Bernstein's inequality explained in Section 8.4.2. Thus, Bennett's inequality gives a tighter upper bound than Bernstein's inequality, although it is slightly more complicated than Bernstein's inequality.