Lisa Yan and Jerry Cain
CS109

Lecture Notes #28
November 16, 2020

Probability Bounds

Based on a chapter by Chris Piech

The following inequalities are useful when you know very little about your distribution, but you would still like to make probabilistic claims. They most often show up in proofs.

Simple Bounds on Expectation

We'll start out with two simple statements about expectation that will be useful for proving the more complex statements below:

$$P(X \geq a) = 1 \qquad \Rightarrow \qquad E[X] \geq a \tag{1}$$

$$P(X \geq Y) = 1 \qquad \Rightarrow \qquad E[X] \geq E[Y] \tag{2}$$

Markov's Inequality

If X is a *non-negative* random variable:

$$P(X \geq a) \leq \frac{E[X]}{a} \qquad \text{for all } a > 0$$

We can prove this statement with our good friend the indicator variable. Let $I = 1$ if $X \geq a$, 0 otherwise. If $X \geq a$, then $\frac{X}{a} \geq 1 = I$. But if $X < a$, then $I = 0$, and since we know X is non-negative, $\frac{X}{a} > 0$. So regardless, $I \leq \frac{X}{a}$. A bit of manipulation of expectation gives us the result:

$$\begin{aligned}
I &\leq \frac{X}{a} \\
E[I] &\leq E\left[\frac{X}{a}\right] && \text{by (\textbf{??}) above} \\
E[I] &\leq \frac{E[X]}{a} && \text{linearity of expectation} \\
P(X \geq a) &\leq \frac{E[X]}{a} && \text{expectation of an indicator variable}
\end{aligned}$$

While the practical use of this is often to put limits on probabilities for distributions whose expectation is known, the intuition for why it is true is a bit clearer if you see it as limiting the expectation given the probability:

$$E[X] \geq a \cdot P(X \geq a)$$

If we hold the probability that $X \geq a$ fixed, and we know that $X \geq 0$, but we want to make the expectation as small as possible, the best we can do is push all the probability density of X that is greater than a down to a, and all the probability density that is less than a down to 0. If we did that, the expectation would become

$$\begin{aligned}
E[X] &= 0 \cdot P(X = 0) + a \cdot P(X = a) \\
&= 0 \cdot P(X < a) + a \cdot P(X \geq a)
\end{aligned}$$

Markov's inequality tells us this is as low as the expectation can go.

Chebyshev's Inequality

If X is a random variable with $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \qquad \text{for all } k > 0$$

We could prove this using indicator variables as above, but it's easier to prove it using Markov's inequality (code re-use!). Let $Y = (X - \mu)^2$. Then $E[Y] = \mathrm{Var}(X) = \sigma^2$. Applying Markov's inequality to Y with $a = k^2$ gives us

$$\begin{aligned} P(Y \geq k^2) &\leq \frac{E[Y]}{k^2} \\ P((X - \mu)^2 \geq k^2) &\leq \frac{\mathrm{Var}(X)}{k^2} \\ P(|X - \mu| \geq k) &\leq \frac{\sigma^2}{k^2} \end{aligned}$$

The intuition is also similar: if we know the expectation and the probability that X is at least k away from the expectation (in either direction), then minimizing the variance means pushing the probability mass inside the interval $(\mu - k, \mu + k)$ all inward to μ, and pushing the probability mass outside that interval inward to $\mu \pm k$. So the smallest variance possible is

$$\begin{aligned} \sigma^2 = E[(X - \mu)^2] &\geq 0^2 \cdot P(|X - \mu| = 0) + k^2 \cdot P(|X - \mu| = k) \\ &= k^2 \cdot P(|X - \mu| \geq k) \end{aligned}$$

There's also a pair of "one-sided" inequalities named after Chebyshev, which say that

$$\begin{aligned} P(X \geq \mu + a) &\leq \frac{\sigma^2}{\sigma^2 + a^2} \\ P(X \leq \mu - a) &\leq \frac{\sigma^2}{\sigma^2 + a^2} \end{aligned}$$

Example 1

Problem: An example has a mean score of 82.2 out of 120 and a standard deviation of 18.5 (variance $\sigma^2 = 342.25$). What's highest possible fraction of students that scored at least 100?

Solution: Assuming scores are nonnegative, Markov's inequality tells us that

$$P(X \geq 100) \leq \frac{E[X]}{100} = 0.822$$

The one-sided Chebyshev's inequality tells us

$$P(X \geq 100 = 82.2 + 17.8) \leq \frac{\sigma^2}{\sigma^2 + 17.8^2} = \frac{342.25}{342.25 + 17.8^2} \approx 0.519$$

We can see that Chebyshev's inequality gives us a tighter bound. This is because we know the variance in addition to the mean.

Still, both are very loose bounds; the actual fraction who got greater than 100 on this particular exam was 0.109.

Problem: What's the highest possible fraction of students that scored more than two standard deviations away from the mean?

Solution: Chebyshev's (two-sided) inequality shows us that

$$P(|X - \mu| > 2\sigma) \leq \frac{\sigma^2}{\sigma^2 + 4\sigma^2} = 0.2$$

In fact, the real fraction of students who were more than two standard deviations from the mean (in either direction) was 0.072.

Jensen's Inequality

If X is a random variable and $f(x)$ is a **convex function** (that is, $f''(x) \geq 0$ for all x), then **Jensen's inequality** says that

$$E[f(X)] \geq f(E[X])$$

A convex function is, roughly speaking, "bowl-shaped", curving upwards. So one way to remember which way the inequality goes is to set up the simplest possible probability distribution: probability 0.5 of being at a and probability 0.5 of being at b. Which is greater: $f(\frac{a+b}{2})$ or $\frac{f(a)+f(b)}{2}$?

Since f curves upward, $f(\frac{a+b}{2})$ is going to lie below (or at most on) the straight line between $(a, f(a))$ and $(b, f(b))$. The average $\frac{f(a)+f(b)}{2}$ is going to lie on that line at $x = \frac{a+b}{2}$, so $\frac{f(a)+f(b)}{2}$ is greater.

(Note that this isn't a proof of the inequality, which holds for other probability distributions besides this simple one.)

You can also show from this that if f is *concave* ($f''(x) \leq 0$ for all x), then $E[f(X)] \leq f(E[X])$.

Applications: Jensen's inequality is commonly used in optimization problems like maximum likelihood estimation. Maximizing the log-likelihood function $L(\theta)$ explicitly might be difficult, but if we could prove that $L(\theta)$ is concave, then Jensen's inequality applies. We could then instead repeatedly try to maximize this lower bound on the maximum expectation. For more information, look up the EM algorithm or take CS229: Machine Learning!

Law of Large Numbers

Consider IID random variables $X_1, X_2, \ldots, X_n$ such that $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Then for any $\varepsilon > 0$, the **weak law of large numbers** states:

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \varepsilon) = 0$$

The **strong law of large numbers** states:

$$P\left(\lim_{n\to\infty}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \mu\right) = 1$$

Both of these say that the sample mean converges to the true mean as we get more samples. The weak version takes the limit of a probability: as you take more samples, the probability of extreme values of the sample mean converges to zero. The strong version takes the probability of a limit: if you imagine the whole (infinite) sequence of random variables consisting of sample means of each size out to infinity, then the limit of that sequence exists and *is* the true mean μ, with probability 1.