CIS3990-002: Mathematics of Machine Learning Fall 2023

Lecture: Concentration

Date: September 6th, 2023 *Author: Eric Wong*

Attribution. These notes are extremely similar to the beginning lectures of Larry Wasserman's Intermediate Statistics course from CMU (`https://www.stat.cmu.edu/~larry/=stat705/`), with some slight notation tweaks to match the course.

1 Concentration Basics

Recall our goal of generalization:

$$\mathbb{P}\left(R_{\text{emp}}(f, X, Y) - R_{\text{true}}(f) < \epsilon\right) > 1 - \delta$$

where

$$R_{\text{emp}}(f, X, Y) = \frac{1}{N}\sum_i \ell(f(x_i, y_i))$$

and

$$R_{\text{true}}(f) = \mathbb{E}_{x,y}\left[\ell(f(x), y)\right]$$

In other words, we want the empirical average to be close to the mean. This is called *concentration*, i.e. the empirical mean concentrates around the true mean.

1.1 Coin flips

Instead of risk, let's consider a much simpler example. Suppose I toss a fair coin n times, and record $x_i = 1$ if heads and $x_i = 0$ otherwise. Consider the average,

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It is easy to see that $\mathbb{E}[\hat{\mu}_N] = 1/2$. How far away is $\hat{\mu}_N$ from its expectation? For example, if $x_i = 1$ for all N flips, then $\hat{\mu}_N = 1$ and it is very far.

Concentration of measure phenomenon says that $\hat{\mu}_N$ "concentrates" closer to $\mathbb{E}[\hat{\mu}_N]$, i.e.

The average of N i.i.d. variables concentrates within an interval of length roughly $1/\sqrt{N}$ around the mean.

- Intuitively, if the average is far from the expectation, then many independent variables need to work together which is extremely unlikely.
- The concentration result is actually stronger: $\hat{\mu}_N$ has an approximately Normal distribution.
- This result underlies pretty much all of statistics and machine learning.

1.2 Tail inequalities

- Markov's inequality: for positive random variable $x \geq 0$ and $\mathbb{E}[X] = \mu < \infty$ then

$$P(X \geq t) \leq \frac{\mu}{t} = O\left(\frac{1}{t}\right)$$

- Very crude, but no distributional assumption, only non-negativity and finite mean!
- "If mean is small, then it is unlikely to be large."
- Proof: basic probability

$$\mathbb{E}[X] = \int_0^\infty xp(x)dx \geq \int_t^\infty xp(x)dx \geq t\int_t^\infty p(x)dx = t\mathbb{P}(X \geq t)$$

- Chebyshev's inequality: for random variable X with finite variance $V(X) = \sigma^2$, for any $t > 0$ we have

$$\mathbb{P}\left(|X - \mu| \geq t\sigma\right) \leq \frac{1}{t^2} = O\left(\frac{1}{t^2}\right)$$

- Proof: apply Markov's inequality

$$\mathbb{P}\left(|X - \mu| \geq t\sigma\right) = P\left(|X - \mu|^2 \geq t^2\sigma^2\right) \leq \frac{\mathbb{E}[|X - \mu|^2]}{t^2\sigma^2} = \frac{1}{t^2}$$

- With more assumptions (finite variance) we can get a better rate $1/t^2$ instead of $1/t$.

Weak Law of Large Numbers (almost). Returning to $\hat{\mu}_N = \frac{1}{N}\sum_i X_i$ (i.e. the coin flip example), note that this has mean μ and variance σ^2/N. Apply Chebyshev's inequality to $\hat{\mu}_N$ and we get:

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}$$

So, with probability at least 0.99 (i.e. by taking $1/t^2 = 0.01$ for $t = 10$), then the average is within $10\sigma/\sqrt{N}$ of the expectation. This is something called the Weak Law of Large Numbers. The key property is the $\frac{1}{\sqrt{N}}$ behavior, with better refinements having dramatically better constants than 10.

- Chernoff Method: introduce a parameter t and an exponential function to refine the Chebyshev inequality.
- For any $t > 0$, we have that

$$\mathbb{P}\left((X - \mu) \geq u\right) = P\left(\exp(t(X - \mu)) \geq \exp(tu)\right) \leq \frac{\mathbb{E}[\exp(t(X - \mu))]}{\exp(tu)}$$

  by applying Markov's inequality.

- Chernoff's bound:

$$\mathbb{P}\left((X-\mu) \geq u\right) \leq \inf_{0 \leq t \leq b} \frac{\mathbb{E}[\exp(t(X-\mu))]}{\exp(tu)}$$

  where b is such that $\mathbb{E}[\exp(tX)]$ (the moment generating function, or mgf) is finite for all $t \leq b$.

- This can be rewritten as

$$\mathbb{P}\left((X-\mu) \geq u\right) \leq \inf_{0 \leq t \leq b} \exp(-t(u+\mu))\mathbb{E}[\exp(tX)]$$

  which is now in terms of the MGF.

Aside: The moment generating function is called such because it can be used to "generate" all the "moments" (i.e. the expected value of X^t for all integer powers of t). Simply write out the Taylor series as

$$M_X(t) = \mathbb{E}[\exp(tX)] = \mathbb{E}\left[1 + tX + \frac{t^2X^2}{2!} + \dots\right] = 1 + t\mathbb{E}[X] + \frac{t^2\mathbb{E}[X^2]}{2!} + \dots$$

Then differentiate i times with respect to t and set $t = 0$ to get the ith moment (i.e. $\mathbb{E}[X^i]$). Fun fact: the form of the MGF specifies the entire distribution (i.e. if you know the MGF then there is only one density it could be). This proof is a bit more technical and can be found in "An Introduction to Probability Theory and Its Applications, Vol. 2" by Feller using Laplace transform theory.

- MGF of a standard normal $N(0,1)$:

$$m_X(t) = \mathbb{E}[\exp(tX)] = \int \exp(tx)\frac{1}{2\pi}e^{-\frac{1}{2}x^2} = \int \frac{1}{2\pi}e^{tx-\frac{1}{2}x^2}dx$$

- Completing the square gets us

$$\int \frac{1}{2\pi}e^{-\frac{1}{2}x^2+tx-\frac{1}{2}t^2+\frac{1}{2}t^2} = \int \frac{1}{2\pi}e^{-\frac{1}{2}(x-t)^2+\frac{1}{2}t^2}dx = e^{\frac{1}{2}t^2}$$

- Example: Gaussian tail bound. Suppose $X \sim N(\mu, \sigma^2)$. Then, if Z is standard Normal, then $X = \sigma Z + \mu$. Then,

$$\mathbb{E}[\exp(tX)] = E[\exp(t(\sigma Z + \mu))] = E[\exp(t\sigma Z)\exp(t\mu)] = \exp(t\mu)m_Z(t\sigma) = \exp(t\mu + \frac{1}{2}t^2\sigma^2)$$

- To apply Chernoff's bound, we compute the minimum over all t:

$$\inf_{t \geq 0} \exp(-t(u+\mu))\exp(t\mu + \frac{1}{2}t^2\sigma^2) = \inf_{t \geq 0} \exp(-tu + \frac{1}{2}t^2\sigma^2)$$

  which is minimized at $t = \frac{u}{\sigma^2}$

- Plug this in to get

$$\mathbb{P}\left((X-\mu) \geq u\right) \leq \exp(-\frac{u^2}{\sigma^2} + \frac{u^2}{2\sigma^2}) = \exp(-\frac{u^2}{2\sigma^2})$$

- This is a one-sided tail bound. Combining with the other side of the tail bound

$$\mathbb{P}\left(|X-\mu| \geq u\right) \leq 2\exp(-\frac{u^2}{2\sigma^2})$$

- This bound is much tighter than Chebyshevs. For $\hat{\mu} = \frac{1}{N}X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$, we have $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N)$.
- Then, the Gaussian tail bound for this where $u = t\sigma/\sqrt{N}$ is

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq t\sigma/\sqrt{N}\right) \leq 2\exp(-\frac{t^2}{2})$$

- Compare to the WLLN variant from before:

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}$$

Aside: Both bounds say the deviation goes down at $\frac{1}{\sqrt{N}}$. However, Gaussian tail bound goes down with exponentially fast. Previously Chebyshev told us with probaiblity 0.99, the average is within $10\sigma/\sqrt{N}$. With the exponential tail bound, with probabilty 0.99 we have that the average is within

$$\sqrt{2\ln(1/0.005)}\sigma/\sqrt{N} \approx 3.25\sigma/\sqrt{N}$$

More generally, Chebyshev says:

$$|\hat{\mu} - \mu| \leq \frac{\sigma}{\sqrt{n\delta}}$$

whereas Gaussian tails tell us

$$|\hat{\mu} - \mu| \leq \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}$$

where the first is polynomial in δ and the second is logarithmic.