CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 2: September 7

Lecturer: Csaba Szepesvári *Scribes: Shivam Garg*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have **not** been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In the previous lecture, we discussed a statistical framework for studying supervised learning. In this lecture, we focus on the question of how to evaluate an algorithm using finite data, i.e. "given an algorithm, can you say whether it will result in a small loss?" To study this question, we will use concentration of measure.

Let us start by a brief review of our notation. We denote the space of the input by $\mathcal{X}$ and the space of the outputs (or labels) as $\mathcal{Y}$. Let $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be the data generating distribution, and $(X'_1, Y'_1), \dots, (X'_m, Y'_m) \overset{\text{iid}}{\sim} P$ be the test data. We denote the loss function by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. For any function $f : \mathcal{X} \to \mathcal{Y}$, we define its empirical loss as $L_m(f) = \big(\sum_{i=1}^m \ell(f(X'_i), Y'_i)\big)/m$, and its expected loss as $L(f) = \int \ell(f(x), y) P(dx, dy)$. Our goal is to find a function that minimizes $L(f)$, however, all we can compute is $L_m(f)$. Then our question becomes "why and when do we think that $L_m(f)$ is a good measure of the expected performance $L(f)$?"

2.1 Measure concentration

Let $X_1, \dots, X_n \sim P \in \mathcal{M}_1(\mathbb{R})$ be iid random variables whose mean $\mu = \int x P(dx)$ and variance $\sigma^2 = \int (x-\mu)^2 P(dx)$ exist. Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. When is it a good idea to use sample mean as an estimate of the true mean? We know that this sample mean is an unbiased estimator of the true mean, i.e. $\mathbb{E}[\hat{\mu}] = \mu$, and has the variance $\mathbb{E}[(\hat{\mu} - \mu)^2] = \sigma^2/n$. Can we say more?

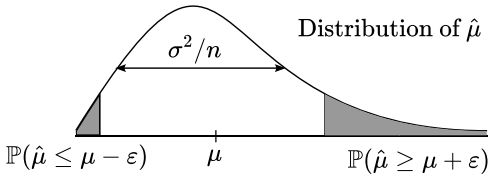

Figure 2.1: Distribution of $\hat{\mu}$ with mean μ and variance σ^2/n. The shaded regions denote the upper and lower tails of the distribution for some $\varepsilon > 0$.

One way is to bound the probability by which $\hat{\mu}$ differs from the true mean μ. For instance, maybe we want the probabilities of both the tail events to be tiny, so that the probability of $\hat{\mu}$ not deviating too much from μ is large:

$$\mathbb{P}(\mu \in [\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon]) \ge 1 - (\text{tiny number}) \ge (\text{big number}),$$

for some small $\varepsilon > 0$.

Central limit theorem (CLT)

One way to go about bounding the deviation between $\hat{\mu}$ and μ is via the central limit theorem. Let $X_1, \dots, X_n$ be a sequence of iid random variables with zero mean, i.e. $\mu = \mathbb{E}[X_i] = 0$. Let $S_n := \sum_{i=1}^n X_i$ denote the sum of these

variables. Let

$$Z_n = \frac{S_n}{\sigma\sqrt{n}} = \frac{S_n}{n}\frac{\sqrt{n}}{\sigma} = \hat{\mu}\frac{\sqrt{n}}{\sigma},$$

denote the normalized sum of the random variables. Then as $n \to \infty$, the central limit theorem says that $P_{Z_n} \to P_Z$, where $Z \sim \mathcal{N}(0,1)$. **(What assumptions does CLT need on the distribution of X_is?)** Recall that for the standard normal distribution, $P_Z(dZ) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$. Then,

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) = \mathbb{P}(\hat{\mu}\sqrt{n}/\sigma \geq \mu + \varepsilon\sqrt{n}/\sigma) \overset{n\to\infty}{\approx} \mathbb{P}(Z \geq \mu + \varepsilon\sqrt{n}/\sigma).$$

For the standard normal distribution, not that for any $u > 0$,

$$\begin{aligned}
\mathbb{P}(Z \geq u) &= \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi u^2}} \int_u^\infty u \cdot e^{-z^2/2} dz \\
&\leq \frac{1}{\sqrt{2\pi u^2}} \int_u^\infty z \cdot e^{-z^2/2} dz \qquad \text{(since the integral is over the set } \{z > u\}) \\
&= \frac{1}{\sqrt{2\pi u^2}} \big[-e^{-z^2/2} \big]_u^\infty = \frac{1}{\sqrt{2\pi u^2}} e^{-u^2/2}.
\end{aligned}$$

Combining the above two results, with $u = \varepsilon\sqrt{n}/\sigma$, gives

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \lesssim \frac{\sigma}{\sqrt{2\pi\varepsilon^2 n}} \exp\left(\frac{-\varepsilon^2 n}{2\sigma^2}\right).$$

Note that this whole argument is not very rigorous, because in order to apply CLT, we need $n \to \infty$, and the above result has a finite n. However, the difference between the cumulative density function of Z_n and Z is not too much. (From Berry-Esséen theorem, this difference is $\mathcal{O}(\mathbb{E}|X_1|^3/\sqrt{n})$.)

Subgaussianity

We present some useful inequalities for bounding the deviation of a random variables and then discuss how a property called subgaussianity can give very fast decays on tail events. The first is Markov's inequality. For any random variable X and $\varepsilon > 0$,

$$\mathbb{P}(|X| \geq \varepsilon) \leq \mathbb{E}|X|/\varepsilon.$$

(If the first moment doesn't exist, then this inequality becomes vacuous.) If the second moment exists, then one can derive, using Markov's inequality, the following (known as Chebyshev's inequality):

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \mathbb{V}[X]/\varepsilon^2.$$

In general, if the pth moment exists, one can obtain $\mathbb{P}(|X - \mu|^p > \varepsilon) \geq \mathbb{E}[|X - \mu|^p]/\varepsilon^p$. (It is further possible to optimize for the value of $p \in \mathbb{N}$ to get the strongest possible inequality, subject to the existence of the pth moment.) But we can do something much simpler. Using Markov's inequality we can obtain the following. For a random variable X with $\mu = \mathbb{E}[X]$ and $\varepsilon > 0$,

$$\mathbb{P}(X \geq \mu + \varepsilon) = \mathbb{P}(\lambda(X - \mu) \geq \lambda\varepsilon) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda\varepsilon}) \leq \mathbb{E}[e^{\lambda(X-\mu)}]/e^{\lambda\varepsilon}, \qquad \text{for all } \lambda > 0.$$

Thus, if we can bound the term $\mathbb{E}[e^{\lambda(X-\mu)}]$, then we can get an exponential decay (which is much faster than what we could obtain by optimizing p in the previous inequalities) on the upper tail event.

Definition 2.1 (Subgaussianity)**.** A random variable X is called σ-subgaussian, if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}$.

Then by the subgaussianity assumption, we get

$$\mathbb{P}(X \geq \mu + \varepsilon) \leq e^{\lambda^2\sigma^2/2 - \lambda\varepsilon} = e^{-\varepsilon^2/(2\sigma^2)}.$$

(Random remark: If the loss ℓ is bounded, then the random variables $L(f_n)$ is also bounded, which means that $L(f_n)$ would be a subgaussian random variable; refer the next lecture.)

2.2 Bibliography