EN.601.776 Modern Topics in Generalization Spring 2026

Lecture 1

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

1 Eliciting Statistics with Losses

Say we're training a model to predict average risk of developing lung cancer within 6 years based on features like age, smoking status, family history, carcinogen exposure, etc. We have access to a dataset D of (x_i, y_i) pairs, where x_i is a vector of features and y_i is in $\{0, 1\}$, indicating whether patient i developed lung cancer within 6 years. We want to train a model h to predict this risk.

Claim 1.1. *The loss function $\ell(h(x), y) = (h(x) - y)^2$ is minimized be the true mean $\bar{y}_x = \mathbb{E}_{(x',y)\sim D}[y \mid x = x']$.*

Proof. Let $D_{y|x}$ be the conditional distribution of y given x. Then

$$\begin{aligned}\mathbb{E}_{y\sim D_{y|x}}[(h(x) - y)^2] &= \mathbb{E}_{y\sim D_{y|x}}[y^2 - 2yh(x) + h(x)^2] \\ &= \mathbb{E}_{y\sim D_{y|x}}[y^2] - 2h(x)\mathbb{E}_{y\sim D_{y|x}}[y] + h(x)^2\end{aligned}$$

We'll differentiate with respect to $h(x)$ and set the derivative to 0 to find the minimum.

$$\begin{aligned}\frac{d}{dh(x)}\mathbb{E}_{y\sim D_{y|x}}[(h(x) - y)^2] &= -2\mathbb{E}_{y\sim D_{y|x}}[y] + 2h(x) \\ &= 0 \\ \implies h(x) &= \mathbb{E}_{y\sim D_{y|x}}[y]\end{aligned}$$

□

Now say we want to predict income based on some set of features, but we want to be robust to outliers since we know Jeff Bezos might be included in our dataset, so we'd like to predict the median income for a given set of features rather than the mean.

Claim 1.2. *The loss function $\ell(h(x), y) = |h(x) - y|$ is minimized be a median ($\hat{y}_x$ such that $\Pr_{y\sim D_{y|x}}[y \leq \hat{y}_x] = \Pr_{y\sim D_{y|x}}[y \geq \hat{y}_x]$).*

Proof. For a given x, we want to find $h(x)$ that minimizes

$$\mathop{\mathbb{E}}_{y\sim D_{y|x}}[|h(x) - y|] = \int_{-\infty}^{\infty} |h(x) - y| D_{y|x}(y)dy.$$

$$
\begin{aligned}
\frac{\partial}{\partial h(x)} \int_{-\infty}^{\infty} |h(x) - y| D_{y|x}(y) dy &= \frac{\partial}{\partial h(x)} \left(\int_{-\infty}^{h(x)} |h(x) - y| D_{y|x}(y) dy + \int_{h(x)}^{\infty} |h(x) - y| D_{y|x}(y) dy \right) \\
&= \int_{-\infty}^{h(x)} \frac{\partial}{\partial h(x)} (h(x) - y) D_{y|x}(y) dy + \int_{h(x)}^{\infty} \frac{\partial}{\partial h(x)} (y - h(x)) D_{y|x}(y) dy \\
&= \int_{-\infty}^{h(x)} D_{y|x}(y) dy - \int_{h(x)}^{\infty} D_{y|x}(y) dy \\
&= \Pr_{y \sim D_{y|x}} [y \leq h(x)] - \Pr_{y \sim D_{y|x}} [y \geq h(x)] \\
&= 0 \implies h(x) \text{ is a median}
\end{aligned}
$$

□

Finally, we now want to predict what number a hand-written digit corresponds to, but our dataset is human annotated. We expect that for any given digit, there will be some incorrect labels, but that the modal label will probably be correct.

Claim 1.3. *The loss function $\ell(h(x), y) = \mathbb{1}[h(x) \neq y]$ is minimized by the modal label ($\hat{y}_x$ such that $\Pr_{y \sim D_{y|x}}[y = \hat{y}_x]$ is maximized).*

Proof. For a given x, $\mathbb{E}_{y \sim D_{y|x}}[\mathbb{1}[h(x) \neq y]] = 1 - \Pr_{y \sim D_{y|x}}[h(x) = y]$. □

2 Introduction

We'll now build up some very helpful techniques from probability theory to bound the probability that $R_S(h)$ deviates from its expectation $R_D(h)$ by more than ε, as a function of m, the size of the sample S. We're not yet concerned with any learning algorithm or how it produced h given training data S. We just want to know, if we're given a model h and a test set $S \sim_{i.i.d.} D^m$, what is the probability that $R_S(h)$ deviates from $R_D(h)$ by more than ε, and how does that probability change as a function of m?

Theorem 2.1 (Markov's Inequality)**.** *Let X be a non-negative random variable. Then for any $a > 0$,*

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Let p denote the PDF of X. Since X is non-negative, we have

$$\begin{aligned}
\mathbb{E}[X] &= \int_{v=0}^{\infty} v \cdot p(v) dv && \text{by definition} \\
&= \int_{v=0}^{a} v \cdot p(v) dv + \int_{a}^{\infty} v \cdot p(v) dv \\
&\geq \int_{v=a}^{\infty} v \cdot p(v) dv && \text{by non-negativity of } X \\
&\geq \int_{v=a}^{\infty} a \cdot p(v) dv && v \geq a \text{ for } v \in [a, \infty] \\
&= a \int_{v=a}^{\infty} p(v) dv && a \text{ is a constant} \\
&= a \cdot \Pr[X \geq a]
\end{aligned}$$

□

We'll apply Markov's inequality to the r.v. $X = (R_S(h) - R_D(h))^2$ with $a = \varepsilon^2$. We'll need to assume we have bounded losses for these arguments, so let's assume $\ell(h(x), y) \in [0, 1]$. Then

$$\begin{aligned}
\Pr_S[|R_S(h) - R(h)| \geq \varepsilon] &= \Pr_S[(R_S(h) - R(h))^2 \geq \varepsilon^2] \\
&\leq \frac{\mathbb{E}[(R_S(h) - R_D(h))^2]}{\varepsilon^2} \\
&= \frac{\mathsf{Var}(R_S(h))}{\varepsilon^2} && \text{by def. of } \mathsf{Var} \\
&= \frac{\mathsf{Var}(\frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i))}{\varepsilon^2} && \text{unpack } R_S(h) \\
&= \frac{\mathsf{Var}(\sum_{i=1}^{m} \ell(h(x_i), y_i))}{m^2\varepsilon^2} && \mathsf{Var}(cX) = c^2\mathsf{Var}(X) \\
&= \frac{\sum_{i=1}^{m} \mathsf{Var}(\ell(h(x_i), y_i))}{m^2\varepsilon^2} && \mathsf{Var}(X_1 + X_2) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) \\
&= \frac{m\mathsf{Var}(\ell(h(x), y))}{m^2\varepsilon^2} && \mathsf{Var}(\ell(h(x_i), y_i)) = \mathsf{Var}(\ell(h(x_j), y_j)) \\
&= \frac{\mathsf{Var}(\ell(h(x), y))}{m\varepsilon^2} \\
&\leq \frac{1}{4m\varepsilon^2} && \text{losses } \in [0, 1] \implies \mathsf{Var} \leq \frac{1}{4}
\end{aligned}$$

So if we want $\Pr_S[|R_S(h) - R(h)| \geq \varepsilon] < \delta$, we can take $m > \frac{1}{4\varepsilon^2\delta}$. As an aside, we have just proven Chebyshev's inequality along the way.

Theorem 2.2 (Chebyshev's Inequality)**.** *Let X be a random variable with non-zero variance $\sigma^2 = \mathsf{Var}(X)$. Then for any $\lambda > 0$*

$$\Pr[|X - \mathbb{E}[X]| \geq \lambda\sigma] \leq \frac{1}{\lambda^2}.$$

Great! So now we have some guarantee that, so long as we take our sample large enough the empirical risk we estimate for our model using a test set will be close to the population risk! But we generally would like really high probability guarantees, which means we want δ to be very, very small. Here we need a sample of size linear in $1/\delta$. We can do much much better by utilizing the fact that our sample is i.i.d., we'll just need some appropriate tail bounds.

Theorem 2.3 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_m$ be independent, bounded random variables with $X_i \in [a_i, b_i]$. Let $S_m = \sum_{i=1}^m X_i$. Then*

$$\Pr_{X_1, X_2, \ldots, X_m}[S_m \geq \mathbb{E}[S_m] + t] \leq e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}.$$

Note this also implies

$$\Pr_{S \sim D^m}[|R_S(h) - R_D(h)| \geq t/m] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

and so for losses bounded in $[0, 1]$, we have

$$\Pr_{S \sim D^m}[|R_S(h) - R_D(h)| \geq t] \leq 2e^{-2t^2 m}$$

We'll begin by proving the inequality assuming the following lemma, which we won't prove in class, but we'll add to the notes for completeness.

Lemma 2.4 (Hoeffding's Lemma)**.** *Let X be a random variable such that $X \in [a, b]$. Then for any $\lambda > 0$,*

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

Proof. (Hoeffding's Inequality) From Markov's inequality, we know that for all $\lambda, t > 0$,

$$\begin{aligned}
\Pr[S_m - \mathbb{E}[S_m] \geq t] &= \Pr[e^{\lambda(S_m - \mathbb{E}[S_m])} \geq e^{\lambda t}] \\
&\leq \frac{\mathbb{E}[e^{\lambda(S_m - \mathbb{E}[S_m])}]}{e^{\lambda t}} && \text{Markov's inequality} \\
&= \frac{\mathbb{E}[e^{\lambda(\sum_{i=1}^m X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} && \text{def of } S_m \text{ and linearity of } \mathbb{E} \\
&= \frac{\mathbb{E}[\prod_{i=1}^m e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} \\
&= \frac{\prod_{i=1}^m \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} && \text{Independence of } X_i\text{s} \\
&\leq \frac{\prod_{i=1}^m e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}}{e^{\lambda t}} && \text{Hoeffding's lemma}
\end{aligned}$$

We showed this is true for all $\lambda > 0$, so in particular it must be true for $\lambda = \frac{4t}{\sum_{i=1}^m (b_i - a_i)^2}$.

Then we have

$$\begin{aligned}
\Pr[S_m - \mathbb{E}[S_m] \geq t] &\leq \frac{\prod_{i=1}^m e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}}{e^{\lambda t}} \\
&= \frac{e^{\frac{\lambda^2}{8} \sum_{i=1}^m (b_i - a_i)^2}}{e^{\lambda t}} \\
&= e^{\frac{\lambda t}{2} - \lambda t} \\
&= e^{-\frac{\lambda t}{2}} \\
&= e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}
\end{aligned}$$

□

Lemma 2.5 (Hoeffding's Lemma). *Let X be a random variable such that $X \in [a, b]$. Then for any $\lambda > 0$,*

$$\mathbb{E}[e^{(\lambda X - \mathbb{E}[X])}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

Proof. We first define a new random variable $Z = X - \mathbb{E}[X]$ and note that $Z \in [c, d]$ for $l = a - \mathbb{E}[X], u = b - \mathbb{E}[X]$ (and so $b - a = u - l$).

By convexity of exp, we have that for all $z \in [c, d]$

$$e^{\lambda z} \leq \frac{u - z}{u - l} e^{\lambda l} + \frac{z - l}{u - c} e^{\lambda d}.$$

It follows that

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &\leq \mathbb{E}\left[\frac{u - Z}{u - l}\right] e^{\lambda l} + \mathbb{E}\left[\frac{Z - l}{u - l}\right] e^{\lambda u} \\
&= \frac{u e^{\lambda l} - l e^{\lambda u}}{u - l} & \mathbb{E}[Z] = 0.
\end{aligned}$$

Where do we go now? If we could show that $\mathbb{E}[e^{\lambda Z}] \leq e^{F(\lambda(u-l))}$, for some function F, and then bound $F(x) \leq \frac{x^2}{8}$, we'd be set. So let's try to massage that last equality into the right form. We want to find F such that

$$e^{F(\lambda(u-l))} = \frac{u e^{\lambda l} - l e^{\lambda u}}{u - l}$$

Note that $\lambda l = \frac{\lambda(u-l)l}{u-l}$ and $\lambda u = \frac{\lambda(u-l)u}{u-l}$. So writing $x = \lambda(u - l)$, our goal is to find F

such that

$$\begin{aligned}
e^{F(x)} &= \frac{ue^{\frac{xl}{u-l}} - le^{\frac{xu}{u-l}}}{u-l} \\
&= e^{\frac{xl}{u-l}}\left(\frac{u - le^{\frac{x(u-l)}{u-l}}}{u-l}\right) && \text{pull out } e^{\frac{xl}{u-l}} \\
&= e^{\frac{xl}{u-l}}\left(\frac{u-l+l-le^{x}}{u-l}\right) && \text{add } 0 \\
&= e^{\frac{xl}{u-l}}\left(1+\frac{l-le^{x}}{u-l}\right)
\end{aligned}$$

So

$$\begin{aligned}
F(x) &= \ln\left(e^{\frac{xl}{u-l}}(1+\tfrac{l-le^x}{u-l})\right) \\
&= \ln(e^{\frac{xl}{u-l}}) + \ln(1+\tfrac{l-le^x}{u-l}) \\
&= \frac{xl}{u-l} + \ln(1+\tfrac{l(1-e^x)}{u-l})
\end{aligned}$$

How do we show this is less than $\frac{\lambda^2(b-a)^2}{8}$? We'll apply Taylor's theorem to $F(x)$ around 0.

Theorem 2.6 (Taylor (specific to our applications))**.** *If a real-valued function F is twice-differentiable at $x=0$, then there exists some $\gamma \in [0,1]$ such that*

$$F(x) = F(0) + xF'(0) + \tfrac{x^2}{2}F''(\gamma x)$$

We have $F(0) = 0$. What about $F'(0)$?

$$\begin{aligned}
F'(x) &= \frac{l}{u-l} + \frac{\frac{df(x)}{dx}}{f(x)} && \text{for } f(x) = 1+\tfrac{l(1-e^x)}{u-l} \\
\frac{df(x)}{dx} &= \frac{-le^x}{u-l} \\
\text{so } F'(0) &= \frac{l}{u-l} - \frac{\frac{l}{u-l}}{1} = 0.
\end{aligned}$$

One more!

$$\begin{aligned}
F''(x) &= \frac{d}{dx}\left(\frac{l}{u-l} + \frac{\frac{-le^x}{u-l}}{1+\frac{l(1-e^x)}{u-l}}\right)\\
&= \frac{d}{dx}\left(\frac{\frac{-le^x}{u-l}}{1+\frac{l(1-e^x)}{u-l}}\right)\\
&= \frac{d}{dx}\left(\frac{-le^x}{u-l+l(1-e^x)}\right)\\
&= \frac{d}{dx}\left(\frac{-le^x}{u-le^x}\right)\\
&= \frac{-lue^x}{(u-le^x)^2}
\end{aligned}$$

From the AMGM inequality, we know that $-lue^x \leq \frac{(u-le^x)^2}{4}$, so we have $F''(x) \leq 1/4$ for all x! Then Taylor's theorem tells us that there's some $\gamma \in [0,1]$ such that

$$\begin{aligned}
F(x) &= F(0) + xF'(0) + \tfrac{x^2}{2}F''(\gamma x)\\
&\leq \frac{x^2}{8}
\end{aligned}$$

and we're done!

Putting it all together we have

$$\mathbb{E}[e^{\lambda Z}] \leq e^{F(\lambda(u-l)} = e^{F(\lambda(b-c))} \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

□