DS 102 Discussion 3

Monday, February 10, 2020

1. What does it mean for the score function of a decision rule to be calibrated? Calibration is a way of assessing the quality of your decision rule when you're interested in the probabilities that a rule assigns to decisions, not just the decisions themselves. For example, when you check the weather and it says there's a 20% chance of rain, you expect that 20% to be a meaningful number. That is, 20% of the time it says there is a 20% chance of rain, it actually rains. From weather, election, and sports forecasts to medical prognoses, it's usually the quality of these predicted probabilities that we care most about.

   To define calibration, we consider decision rules $\delta(X)$ that threshold score functions $R(X)$ that take on values between 0 and 1:

   $$\delta(X) = \mathbb{1}[R(X) \geq \gamma].$$

   As usual, we let $Y \in \{0, 1\}$ denote the true binary label corresponding to X. We call $R(X)$ **calibrated** when, for all values $r \in [0, 1]$, we have

   $$\mathbb{P}(Y = 1 \mid R(X) = r) = r.$$

   Decision rules can achieve high accuracy (or any desirable metric from the confusion matrix) with wildly uncalibrated score functions. If the same weather forecast just rounded its predicted probability of rain to 0% or 100%, as a decision rule for whether it rains, it would achieve the same confusion matrix as before. However, its predicted probabilities would be uncalibrated and therefore useless.

   We'll look more closely at what it means (and doesn't mean) for a decision rule to be calibrated.

   (a) Recall that the conditional expectation $\mathbb{E}[Y|X]$ is the decision rule that minimizes the Bayes risk for squared-error loss. Suppose we use the conditional expectation as our score function:

   $$R(X) = \mathbb{E}[Y|X]$$

   Is the score function calibrated?

   Solution: Yes. For $r \in [0, 1]]$, we have

   $$\begin{aligned}\mathbb{P}_{X,Y}(Y = 1 \mid R(X) = r) &= \mathbb{P}_{X,Y}(Y = 1 \mid \mathbb{E}[Y|X] = r) \\ &= \mathbb{P}_{X,Y}(Y = 1 \mid \mathbb{P}(Y = 1|X) = r) \\ &= r.\end{aligned}$$

   (b) In practice, we don't know $\mathbb{P}_{X,Y}(Y = 1 \mid R(X) = r)$. However, we can estimate it given a dataset. Consider the following table of data X, labels Y, and scores $R(X)$. Is this score function calibrated on this dataset? Does it seem like a good score function for a decision rule, in terms of its false positive and false negative rates?

X	Y	$R(X)$
0	0	1/2
0	0	1/2
1	1	1/2
1	1	1/2

Solution: Yes, the simple score function $R(X) = 1/2$ is calibrated: we have $\mathbb{P}(Y = 1 \mid R(X) = 1/2) = 1/2$. However, it has high false positive and false negative rates of $1/2$. It is often important to have a calibrated score function, but on its own, it is not sufficient for a useful decision rule.

(c) For any particular value of $X = x$, do we have

$$\mathbb{P}(Y = 1 \mid R(X) = 1/2, X = x) = 1/2?$$

Solution: No. For $X = 0$, we have $\mathbb{P}(Y = 1 \mid R(X) = 1/2, X = 0) = 0$, and for $X = 1$, we have $\mathbb{P}(Y = 1 \mid R(X) = 1/2, X = 1) = 1$. That is, for a calibrated score function, it is **not** the case that for any particular instance with a score of r, the probability that that instance has $Y = 1$ is necessarily r.

(d) Let's return to the conditional expectation $R(X) = \mathbb{E}[Y|X]$. For any particular value of $X = x$, do we have

$$\mathbb{P}_{X,Y}(Y = 1 \mid R(X) = r, X = x) = r?$$

Solution: Yes, by similar proof as Part (a). By the same reasoning, for any set S we might be interested in, we have

$$\mathbb{P}_{X,Y}(Y = 1 \mid R(X) = r, X \in S) = r.$$

2. **Linear Regression and Ordinary Least Squares:** For this question, we will review the Ordinary Least Squares (OLS) estimator for linear regression. In addition to deriving the OLS estimator, we will add on a new probabilistic interpretation by introducing the idea of an **empirical distribution**.

   Setup: We observe n data points

   $$x^{(1)}, ..., x^{(n)} \in \mathbb{R}^d,$$

   and corresponding outputs

   $$y^{(1)}, ..., y^{(n)} \in \mathbb{R}.$$

We assume that

$$y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \epsilon^{(i)},$$

where β^* is some true parameter that we want to estimate, and $\epsilon^{(i)}$ represents some random error.

The OLS estimator is the estimator $\hat{\beta}$ that minimizes the *sum of squared residuals*, where each *residual* is the difference between each true $y^{(i)}$ and the estimated $\langle \hat{\beta}, x^{(i)} \rangle$. Note that the random errors $\epsilon^{(i)}$ are not taken into account in the residuals – this means that the OLS estimator might be bad if the $\epsilon^{(i)}$ are not zero-mean.

(a) Write the minimization problem that gives us the OLS estimator.

Solution: Let $\hat{\beta}$ be the OLS estimator. Then $\hat{\beta}$ is given by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (y^{(i)} - \langle \beta, x^{(i)} \rangle)^2$$

(b) Solve the minimization problem by taking the derivative with respect to β and setting that equal to 0. Recall from the matrix cookbook that the derivative with respect to a vector β of $\langle \beta, x \rangle$ is $\frac{\partial}{\partial \beta} \langle \beta, x \rangle = x$. Also recall that $\langle \beta, x \rangle = \beta^T x = x^T \beta$.

Solution: Note to GSIs: feel free to skip some of the algebra in this derivation to get to part (c).

$$\begin{aligned} \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y^{(i)} - \langle \beta, x^{(i)} \rangle)^2 &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y^{(i)})^2 - 2y^{(i)} \langle \beta, x^{(i)} \rangle + \langle \beta, x^{(i)} \rangle^2 \\ &= \sum_{i=1}^{n} -2y^{(i)} x^{(i)} + 2x^{(i)} (x^{(i)})^T \beta \end{aligned}$$

Setting this equal to 0 to find the minimum:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} -2y^{(i)}x^{(i)} + 2x^{(i)}(x^{(i)})^T\hat{\beta} \\
\sum_{i=1}^{n} x^{(i)}(x^{(i)})^T\hat{\beta} &= \sum_{i=1}^{n} y^{(i)}x^{(i)} \\
n\left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^T\right)\hat{\beta} &= \sum_{i=1}^{n} y^{(i)}x^{(i)} \\
\hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y^{(i)}x^{(i)}\right)
\end{aligned}$$

(c) **Sample means and empirical distributions:** Suppose we have some distribution p, and we have n random samples from that distribution, $x^{(1)}, ..., x^{(n)} \sim p$. Then the **sample mean** is defined as

$$\frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$

Another way to describe the sample mean is an **expectation over an empirical distribution**. Specifically, if $\hat{p}_n$ is a distribution of a random variable x where x is equal to each of the samples $x^{(i)}$ with equal probability $\frac{1}{n}$, then

$$E_{x\sim\hat{p}}[x] = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$

We call $\hat{p}_n$ an **empirical distribution** of the original distribution p.

Let $\hat{p}_n$ be the empirical distribution over the data points $(x^{(1)}, y^{(1)}), (x^{(n)}, y^{(n)})$ of the underlying distribution that the data points were truly drawn from. That is, for a random variable $(x, y) \sim \hat{p}_n$, $(x, y) = (x^{(i)}, y^{(i)})$ with probability $\frac{1}{n}$. Using the solution to part (b), write the OLS estimator $\hat{\beta}$ in terms of expectations over $\hat{p}_n$.

Solution: In the formula from part (b), we can simply replace every instance of $\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}, y^{(i)})$ with $\mathbb{E}_{(x,y)\sim\hat{p}}[f(x, y)]$.

$$\hat{\beta} = \left(\mathbb{E}_{(x,y)\sim\hat{p}}[xx^T]\right)^{-1}\left(\mathbb{E}_{(x,y)\sim\hat{p}}[xy]\right)$$