Additional notes for Lecture on Monday Oct 5

October 2, 2020

Today we will do three derivations. The derivations are kind of involved and subtle, so I am not satisfied with using the slides. We will go over this in class, but here are a set of notes with more detail.

Derivations 1: Showing $\hat{\sigma}_D^2$ is a biased estimator of variance

- **Setup**: Consider the normal Gaussian distribution $\mathcal{N}(\mu, \sigma)$, e.g. $x \sim \mathcal{N}(\mu, \sigma)$ is x sampled from a Gaussian distribution with mean μ and variance σ^2.
- We define a *draw* $\mathcal{D}$ as a collection of samples $\mathcal{D} = \{x_1, ..., x_m\}$ (where $m = |\mathcal{D}|$ is the number of samples in our draw), and each x_i is drawn i.i.d. from $\mathcal{N}(\mu, \sigma)$.
- Previously, we defined two sample-dictated estimates of μ and σ, namely

$$\hat{\mu}_{\mathcal{D}} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \hat{\sigma}^2_{\mathcal{D}} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu}_D)$$

  We had shown that $\hat{\mu}_{\mathcal{D}}$ is unbiased, e.g. $\mathbb{E}[\hat{\mu}_{\mathcal{D}}] = \mu$.
- At this point, it's worth thinking about what we are taking the expectation with respect to. Is it $\mathcal{D}$? is it x_i? Turns out it's both. Think of it like collecting data for an experiment. On Monday, you collect 10 data readings. On Tuesday, you collect 10 more. (All on the same experiment.) Then knowing $\mathcal{D}$ tells you what day you collected the data, and knowing $x_i \in \mathcal{D}$ tells you what the data was that you collected on that day. Both are random.
- Now let's take a look at the biased-ness of $\hat{\sigma}^2_{\mathcal{D}}$:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_{\mathcal{D}}] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu}_{\mathcal{D}})^2\right] \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu}_{\mathcal{D}} + \mu - \mu)^2\right] \\
&= \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[(x_i - \mu)^2]}_{\sigma^2} + \mathbb{E}[(\mu - \hat{\mu}_{\mathcal{D}})^2] + \underbrace{\frac{1}{m}\sum_{i=1}^{m}(2\mathbb{E}[(x_i - \mu)(\mu - \hat{\mu}_{\mathcal{D}})])}_{\text{Term } \star}
\end{aligned}$$

  Hopefully, so far, nothing controversial. Now let's look at term $\star$ more carefully. This is the part where I got stuck in class, because in fact we have some coupling: both x_i and $\hat{\mu}_{\mathcal{D}}$ are random variables, and are *not* independent from each other. So we have to figure out how to split the two.
- In particular, we are going to split the expectation with respect to our two sources of randomness: the draw itself (think "experiment day") and the sample from the draw (think "experiment datapoint on that day")

$$\mathbb{E}[(x_i-\mu)(\mu-\hat{\mu}_{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{x_i\in\mathcal{D}}[(x_i-\mu)(\mu-\hat{\mu}_{\mathcal{D}})|\mathcal{D}]\right]$$

Now, $\hat{\mu}_{\mathcal{D}}$ is constant *given* $\mathcal{D}$! So the inside expectation

$$\mathbb{E}_{x_i\in\mathcal{D}}[(x_i-\mu)(\mu-\hat{\mu}_{\mathcal{D}})|\mathcal{D}] = \left(\underbrace{\mathbb{E}_{x_i\in\mathcal{D}}[x_i|\mathcal{D}]}_{\hat{\mu}_{\mathcal{D}}}-\mu\right)(\mu-\hat{\mu}_{\mathcal{D}}) = -(\hat{\mu}_{\mathcal{D}}-\mu)^2.$$

Overall, this gives us at this point

$$\mathbb{E}[\hat{\sigma}^2_{\mathcal{D}}] = \sigma^2 - \mathbb{E}_{\mathcal{D}}[(\hat{\mu}_{\mathcal{D}}-\mu)^2].$$

- Finally, we need to expand the second term. Namely,

$$\mathbb{E}_{\mathcal{D}}[(\hat{\mu}_{\mathcal{D}}-\mu)^2] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m x_i-\mu\right)^2\right] = \frac{1}{m^2}\underbrace{\mathbb{E}\left[\left(\sum_{i=1}^m(x_i-\mu)\right)^2\right]}_{s(m)}$$

Let's stare at this last term a bit closer:

$$\begin{aligned} s(m) &= \mathbb{E}\left[\left(\sum_{i=1}^m(x_i-\mu)\right)^2\right] \\ &= \mathbb{E}\left[(x_m-\mu)^2+2(x_m-\mu)\left(\sum_{i=1}^{m-1}(x_i-\mu)\right)+\left(\sum_{i=1}^{m-1}(x_i-\mu)\right)^2\right] \\ &= \underbrace{\mathbb{E}[(x_m-\mu)^2]}_{\sigma^2}+2\underbrace{\mathbb{E}\left[(x_m-\mu)\left(\sum_{i=1}^{m-1}(x_i-\mu)\right)\right]}_{\square}+\underbrace{\mathbb{E}\left[\left(\sum_{i=1}^{m-1}(x_i-\mu)\right)^2\right]}_{s(m-1)} \end{aligned}$$

Again, we have an issue where, though each x_i are i.i.d., when they are multiplied to each other, then it's not so easy to decouple. So, we use the same trick as before, which is the "law of conditional expectation"; namely, for any function and any random variable X and Y (which may or may not be independent),

$$\mathbb{E}[f(X,Y)] = \mathbb{E}_X[\mathbb{E}_Y[f(X,Y)|X]].$$

So, let's try to separate each data point from each other, one by one. We start with x_m. Then

$$\square = \mathbb{E}_{x_1,\ldots,x_{m-1}}\left[\mathbb{E}_{x_m}\left[(x_m-\mu)\underbrace{\left(\sum_{i=1}^{m-1}(x_i-\mu)\right)}_{\Delta}\right]|x_1,\ldots,x_{m-1}\right]$$

Well, given $x_1,\ldots,x_{m-1}$, then the Δ term is just a constant! Then $\mathbb{E}[x_i-\mu]=0$, and therefore $\square=0$. We are then left with the recursion

$$s(m) = \sigma^2 + s(m-1) = m\sigma^2.$$

Piecing it all together, we get

$$\mathbb{E}_{\mathcal{D}}[(\hat{\mu}_{\mathcal{D}} - \mu)^2] = \frac{1}{m^2} \cdot m\sigma^2 = \frac{\sigma^2}{m}$$

and overall,

$$\mathbb{E}[\hat{\sigma}^2_{\mathcal{D}}] = \left(1 + \frac{1}{m}\right)\sigma^2.$$

- Followup question: Having access only to $x_i \in \mathcal{D}$, how would I design an *unbiased* estimator of the variance?

Derivation 2: Asymptotic consistency of 1-NN estimator

Ok, now we're going to switch gears almost completely, and talk about the 1-NN classifier.

- Assume we live in a world where datapoints $x \sim \mathcal{X}$ are distributed randomly, and for each datapoint, that datapoint is labeled with probability $p > 0$. (We can debate the validity of the assumption, but it's what we need to make our claim.)
- Suppose we choose some distance function, say $d(x, y)$, and the only thing we impose on this distance function is that it's continuous with respect to its input, $d(x, y) > 0$ if $x \neq y$, and $d(x, x) = 0$. Any norm would work, but other functions would work also.
- Now assume that we have m datapoints $x_i \in \mathcal{D}$ ($|\mathcal{D}| = m$), littered in our feature space, each drawn i.i.d. from $\mathcal{X}$, and with probability p is labeled.
- **Claim** Under this regime, For an *unlabeled* point $\bar{x}$,

$$\mathbf{Pr}\left(\lim_{m\to\infty} \min_{\substack{x_i \in \mathcal{D} \\ \text{labeled}}} d(x_i, \bar{x}) = 0\right) = 1$$

- **Proof.** The proof is by contradiction, e.g. we try to disprove that

$$\mathbf{Pr}\left(\lim_{m\to\infty} \min_{\substack{x_i \in \mathcal{D} \\ \text{labeled}}} d(x_i, \bar{x}) > 0\right) > 0.$$

  In particular, the inner clause basically says that there exists a $\delta > 0$ where

$$\lim_{m\to\infty} \min_{\substack{x_i \in \mathcal{D} \\ \text{labeled}}} d(x_i, \bar{x}) = \delta > 0$$

  with nonzero probability.
- If this were true, that would mean that there is a region around $\bar{x}$, let's define it as

$$R_\delta(\bar{x}) = \{x : d(x, \bar{x}) \leq \delta\}$$

  and $\bar{x} \in R_\delta(\bar{x})$ (since $\delta(\bar{x}, \bar{x}) = 0 < \delta$), but all $x_i \in \mathcal{D}$, $x_i \notin \mathbb{R}_\delta(\bar{x})$ whenever x_i is labeled.
- So, one of two possibilities can happen. One is that, for any $x \in \mathcal{X}$, $\mathbf{Pr}(x \in R_\delta(\bar{x})) = 0$. Then $\bar{x} \in R_\delta(\bar{x})$ is a 0-probability event.
- Therefore, it must be that, for any $x \in \mathcal{X}$, $\mathbf{Pr}(x \in R_\delta(\bar{x})) = \epsilon > 0$. I don't need to know what ϵ is, but I know it's greater than 0.

- Therefore, the probability that a point x_i is not in $R_\delta(\bar{x})$ is $1-\epsilon$. Moreover, the probability that a point x_i is in $R_\delta(\bar{x})$ and is also labeled is $\delta\epsilon$. Finally, this gives

  $$\mathbf{Pr}(\text{any single point is not in } R_\delta(\bar{x}) \text{ or not labeled}) = 1 - \mathbf{Pr}(\text{any single point is in } R_\delta(\bar{x}) \text{ and labeled}) = 1 - \delta\epsilon.$$

- Therefore, since each point x_i is i.i.d., the probability that, after m draws, there are *no points* in $R_\delta(\bar{x})$ that are labeled, is

  $$\mathbf{Pr}(\text{any single point is in } R_\delta(\bar{x}) \text{ and labeled})^m = (1-\delta\epsilon)^m$$

  which, taking the limit $m \to +\infty$, goes to 0.

- Therefore, at the asymptotic limit ($m \to +\infty$), the only positive probability events result in the closest labeled point being *arbitrarily close to* any unlabeled point $\bar{x}$.

Derivation 3: Asymptotic bounds of 1-NN estimator

Ok! Are you tired yet? It's a lot going on. Don't feel like any of that was supposed to be easy. I spent quite a few tries going over all three derivations before I was convinced they were correct. If you need a break, go get a coffee, and go over this part tomorrow. If not, let's roll!

- First, some build up. We went over Bayes risk and Bayes classifier in class last week. We mentioned that in general, getting, or even describing, the Bayes classifier can be difficult. But, in the case of binary classification, at least describing it is not that hard.

- In particular, in binary classification, we may design a 0-1 loss, e.g.

  $$\mathcal{L}(y(x), \hat{y}) = \begin{cases} 1 & \text{if } y(x) \neq \hat{y} \\ 0 \text{ else.} \end{cases}$$

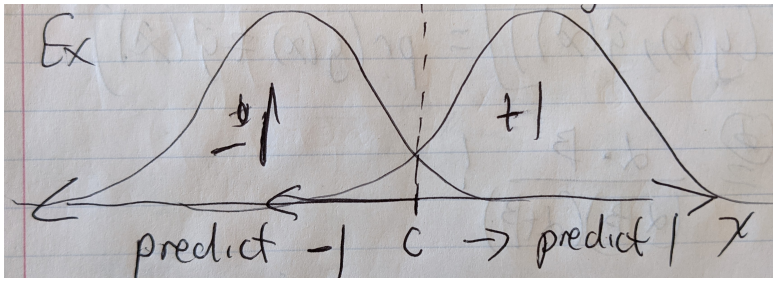
  Then the Bayes risk is

  $$\begin{aligned} \text{Bayes risk } \mathbb{E}[\mathcal{L}(y(x), \hat{y})] &= 1 \cdot \mathbf{Pr}(y(x) \neq \hat{y}) + 0 \cdot \mathbf{Pr}(y(x) = \hat{y}) \\ &= \mathbf{Pr}(y(x) \neq \hat{y}) \end{aligned}$$

- The Bayes estimator, is then (surprise surprise) the estimator that, given x, picks $\hat{y}$ to be the one most likely to equal $y(x)$

  $$\hat{y}_{\text{Bayes}} = \begin{cases} 1 & \text{if } \mathbf{Pr}(y(x)=1) > \mathbf{Pr}(y(x)=-1) \\ -1 & \text{if } \mathbf{Pr}(y(x)=-1) > \mathbf{Pr}(y(x)=1). \end{cases}$$

  Below is a figure showing an example in 1-D. Imagine we have 2 Gaussian distributions, one describing $x|y(x)=1$ and the other describing $x|y(x)=-1$. Then there is some point c where if $x > c$ then $\mathbf{Pr}(y(x)=1) > \mathbf{Pr}(y(x)=-1)$, and if $x < c$, then $\mathbf{Pr}(y(x)=-1) > \mathbf{Pr}(y(x)=1)$.

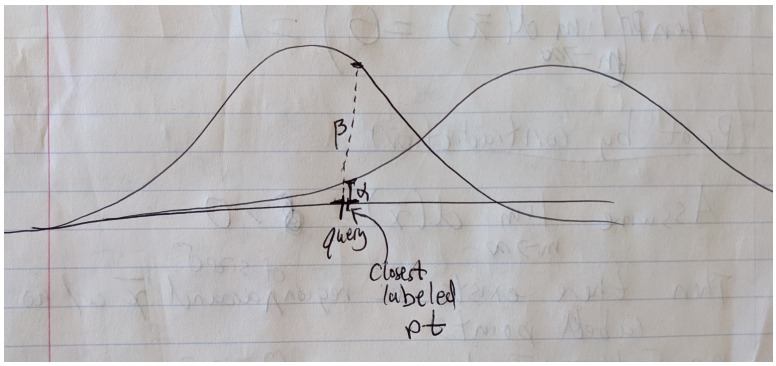


Then, our prediction rule is simply

$$\hat{y}(x) = \begin{cases} 1 & \text{if } x > c \\ -1 & \text{else.} \end{cases}$$

(What happens at $x = c$ is of no concern to us, since it is a 0 probability event.)

- Next, let's take a look at the 1-NN estimator. To get an idea of where we're going with this, take a look at the same 2-Gaussian example, in the figure below.



Now, consider a query point (as labeled) and a nearest labeled point (also labeled). In the limit as query $\to$ nearest label (which we just shows happens as $m \to +\infty$), these two points converge to each other. Then,

$$\mathbf{Pr}(\text{query point label } y(x) = c) = \mathbf{Pr}(\text{labeled point label} y(z_x) = c) = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } c = 1 \\ \frac{\beta}{\alpha+\beta} & \text{if } c = -1. \end{cases}$$

Check your understanding: Does it make sense that, even as $x \to z_x$, that the probability of $\mathbf{Pr}(y(x) = y(z_x)) \neq 1$; and rather, it's not even close?

In particular, what is the probability that $y(z_x) \neq y(x)$? Well, we just use the independence assumption, and get

$$\mathbf{Pr}(y(x) \neq y(z_x)) = \mathbf{Pr}(y(x) = 1)\mathbf{Pr}(y(z_x) = -1) + \mathbf{Pr}(y(x) = -1)\mathbf{Pr}(y(z_x) = 1) = 2\frac{\alpha\beta}{(\alpha+\beta)^2}.$$

- In general, let's define $\eta(x) = \mathbf{Pr}(y(x) = 1)$. Then the Bayes risk of the 1-NN classifier is

$$\text{risk}_{\text{Bayes}}(\hat{y}_{1\text{NN}}) = \mathbf{Pr}(y(x) \neq y(z_x)) = 2\eta(x)(1 - \eta(x)).$$

and as a reminder,

$$\text{risk}_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \min\{\eta(x), 1 - \eta(x)\}.$$

- Now, let's take a step back and look at the whole picture. Literally. Let's take these expectations over query points x. Specifically, define

$$R^* := \mathbb{E}_x[\text{risk}_{\text{Bayes}}(\hat{y}_{\text{Bayes}})], \qquad R_{1\text{NN}} := \mathbb{E}_x[\text{risk}_{\text{Bayes}}(\hat{y}_{1\text{NN}})].$$

Somehow, these constants tell us how good these two classifiers are at minimizing the expected risk. We know that the Bayes classfier is Bayes optimal, so we know that R^* lower bounds this average risk over *any* classifier, but we will prove that anyway. Additionally, we will show how the Bayes classifier risk can upper bound $R_{1\text{NN}}$ as well.

- **Claim**:
$$R^* \leq R_{1\text{NN}} \leq 2R^*(1 - R^*).$$

- **Lower bound proof.** Actually, this just comes from staring at a graphing calculator. For any value $0 \leq \eta \leq 1$, the quantities
$$\min\{\eta, 1 - \eta\} \leq \eta(1 - \eta).$$
Therefore, taking expectations will preserve this inequality.

- **Upper bound proof.** This one requires a bit more finesse. We need 2 tricks. The first trick is to introduce
$$\beta(x) = \min\{\eta(x), 1 - \eta(x)\}.$$
Then $\beta(x)(1 - \beta(x)) = \eta(x)(1 - \eta(x))$, and
$$R^* = \mathbb{E}_x[\beta(x)], \quad R_{1\text{NN}} = \mathbb{E}_x[\beta(x)(1 - \beta(x))].$$
The second trick is to notice that, for any random variable β,
$$\mathbb{E}[\beta^2] \geq (\mathbb{E}[\beta])^2.$$
Why, you ask? Well, let's take a look at what this quantity is:
$$\mathbb{E}[\beta^2] - (\mathbb{E}[\beta])^2 = \text{Var}(\beta) \geq 0.$$
Cute, right?
Overall, this gives
$$\begin{aligned} R_{1\text{NN}} &= 2\mathbb{E}_x[\beta(x)(1 - \beta(x))] \\ &= 2\mathbb{E}_x[\beta(x)] - 2\mathbb{E}_x[\beta(x)^2] \\ &\leq 2\mathbb{E}_x[\beta(x)] - 2(\mathbb{E}_x[\beta(x)])^2 \\ &= 2R^* - 2(R^*)^2 \end{aligned}$$
which gives us our upper bound!

- In the challenge problem, you can investigate whether these bounds are tight. (They are, but it's not obvious *when* they're tight.)

- **Check your understanding.** In what kind of distribution of $\beta(x)$ would you expect the upper bound to be tight? (Hint: when does inequality become equality in the upper bound proof?)

Whew! That's all folks!