Statistical Learning, Fall 2022-23

Class notes 2

Reminder: We denote the training data by $T = (\mathbb{X}_{n\times p}, \mathbb{Y}_{n\times 1})$ to differentiate it from the random variables $X \in \mathbb{R}^p, Y \in \mathbb{R}$.

Decision Theory perspective

We said we are fitting a function $\hat{f}(x)$ that we hope approximates $\mathbb{E}(Y|X = x)$ or simply Y. How should we measure approximation accuracy?

The decision theory perspective calls for defining a *Loss function* $L(Y, f(X))$ for prediction, which defines how close $f(X)$ and Y are for any given function f. For example, for regression a natural choice is the squared loss:

$$L(Y, f(X)) = (Y - f(X))^2.$$

The expected prediction error (EPE) of a prediction function f is:

$$EPE(f) = \mathbb{E}_{X,Y}(Y - f(X))^2 = \int (y - f(x))^2 Pr(x, y).$$

We can write this using the law of iterated expectation:

$$EPE(f) = \mathbb{E}_X \left[\mathbb{E}\left((Y - f(X))^2|X\right)\right],$$

which tells us that the optimal $f(x)$ minimizes the internal expectation for $X = x$.

It is well known that this problem is solved by $\mathbb{E}(Y|X = x) = \arg\min_c \mathbb{E}\left((Y - c)^2|X = x\right)$ Let's confirm it:

$$\begin{aligned}
&\mathbb{E}\left((Y - c)^2 \mid X = x\right) = \mathbb{E}\left[(Y - \mathbb{E}(Y \mid X = x) + \mathbb{E}(Y \mid X = x) - c)^2 \mid X = x\right] = \\
&= \mathbb{E}\left[(Y - E(Y|X = x))^2 \mid X = x\right] + 2\mathbb{E}\left(Y - \mathbb{E}(Y|X = x) \mid X = x\right)\left(\mathbb{E}(Y|X = x) - c\right) + \left(\mathbb{E}(Y|X = x) - c\right)^2 = \\
&= Var(Y \mid X = x) + 0 + (\mathbb{E}(Y \mid X = x) - c)^2 \;\geq\; Var(Y \mid X = x),
\end{aligned}$$

and setting $c = \mathbb{E}(Y \mid X = x)$ clearly attains equality in the last line.

Conclusion: If our prediction loss is squared error, we are trying to learn this conditional expectation functions as the "best possible" prediction model: $f^*(x) = \mathbb{E}(Y \mid X = x)$.

Absolute loss: Similarly, if $L(Y, f(X)) = |Y - f(X)|$, the optimal model is the conditional median $f^*(x) = median(Y \mid X = x)$. Proof: in HW1.

Back to squared loss, with this (not surprising) understanding that we are trying to estimate $f^* = \mathbb{E}(Y \mid X = x)$, the conditional expectation, we can now interpret k-NN as trying to accomplish this directly at point x by:

- Replacing the point $X = x$ by the neighborhood of x in the training data $N_k(x)$
- Replacing the expectation by the average of the k training neighbors

and we can also think of changing k as balancing between the inaccuracies of the two approximations. When we increase k:

- The average is more accurate and stable (we can think of this as reducing variance)
- The neighborhood is bigger and less representative of $X = x$ (increasing the bias)

However with proper asymptotics we can provably get the optimal performance with $k - NN$:
Theorem: If $Pr(x, y)$ is *regular*, $f^*(x)$ is well behaved, $n \to \infty,\ k(n) \to \infty,\ k(n)/n \to 0$, then: $\hat{f}(x) \to f^*(x)\ , \forall x$
Proof: HW1 extra credit (will discuss in class).

For linear regression, we can examine what the optimal linear model looks like:

$$\begin{aligned} \tfrac{d}{d\beta} \quad & \int (y - x\beta)^2 Pr(x, y) = \int -2x(y - x\beta) Pr(x, y) = -2\mathbb{E}(XY) + 2\mathbb{E}(XX^T)\beta \\ \Rightarrow \quad & \beta^* = \left[\mathbb{E}(XX^T)\right]^{-1} \mathbb{E}(XY). \end{aligned}$$

Now we can interpret least squares regression in this context in two ways:

- Estimating each part separately: $\mathbb{E}(XX^T) \approx 1/n \cdot \sum_i x_i x_i^T = 1/n \cdot \mathbb{X}^T\mathbb{X}$, $\mathbb{E}(XY) \approx 1/n \cdot \sum_i x_i y_i = 1/n \cdot \mathbb{X}^T\mathbb{Y}$. And then naturally:

$$\hat{\beta} = \left[\mathbb{X}^T\mathbb{X}\right]^{-1} \mathbb{X}^T\mathbb{Y} \approx \beta^*.$$

- Empirical risk minimization: $\hat{\beta} = \arg\min RSS(\beta)$ with the training set replacing the population.

What about classification?

We can take a similar decision theoretic view of classification, where $Y \in \mathcal{G}$ with $|\mathcal{G}| = K$ classes. We can define $L(Y, g(X))$ via a $K \times K$ matrix of *misclassification loss*:

$$\begin{aligned} L_{kk} &= 0\ , k = 1 \ldots K \quad \text{(no penalty for correct prediction)} \\ L_{jk} &> 0\ , j \neq k \quad \text{(cost of each kind of error)} \end{aligned}$$

Now we can define:

$$EPE(g) = E_{X,Y}L(Y, g(X)) = E_X \left[\sum_{k=1}^{K} L_{k,g(X)} Pr(Y = \mathcal{G}_k \mid X) \right],$$

which again we can minimize at each point x.
The obvious simplest loss function is $0-1$ loss: $L_{jk} = 1,\ \forall j \neq k$. Then we can write the optimizer simply as:

$$g^*(x) = \arg\min_{l \leq K} \sum_k L_{k,l} Pr(Y = \mathcal{G}_k | X = x) = \arg\min_l 1 - Pr(Y = \mathcal{G}_l | X = x) = \arg\max_l Pr(Y = \mathcal{G}_l | X = x).$$

This is the well known *Bayes classifier* which chooses the most likely outcome at every point $X = x$.

In this case again we can think of k-NN as doing the empirical approximations, choosing the "Bayes" decision based on the training neighborhood.

Decision theoretic connection between 2-class classification and regression

For 2-class encode as usual $\mathcal{G} = \{0, 1\}$. What minimizes EPE with squared loss for this problem?

$$f^*(x) = \mathbb{E}(Y \mid X = x) = Pr(Y = 1 \mid X = x),$$

so we can think of the regression solution as estimating the probability, which is what we need for Bayes prediction, which is simply $\mathbb{I}\{\mathbb{E}(Y \mid X = x) > 0.5\}$ in this case.

Curse of dimensionality (COD)

COD refers to the basic, intuitive(?) fact that high dimensional spaces are a lot harder to fill up with points than lower dimensional spaces, and in most senses we require number of points that is exponential in the dimension, to fill a high dimensional spaces.

One way of demonstrating this phenomenon is by considering the relation between radius of a region (=distance to closest neighbor) and which portion of the space it occupies. For example, consider a case where our data space for x is a p-dimensional cube $\mathcal{X} = [0, 1]^p$ (can also be a ball). Then the volume of a cube of side r is r^p , so for example a cube which contains $1/10$ of the volume will be of radius $10^{-1/p} \overset{p \to \infty}{\longrightarrow} 1$. Already in dimension 10, to catch 0.1 of the volume we need $r^{10} = 0.1 \Rightarrow r = 0.8$ or for $0.01 : r = 0.63$. Hence with 100 observations in 10 dimensions we already expect the closest neighbor to be very far away!

More concrete example: imagine now we have n observations distributed uniformly in a unit $\mathbb{R}^P$ ball. We have a prediction point in the origin, and want to consider the distribution of the distance to its nearest neighbor. Median distance to closest neighbor:

$$(1 - 0.5^{1/n})^{1/p} \text{ , (Proof: HW1)},$$

which means that for $p = 20, n = 5000$ the median distance is 0.64 (two thirds of way to the edge), with $p = 100, n = 10^7$ it is 0.85!

Specific conclusion: in high dimension, the closest neighbors are very far away, unless we have an exponential amount of data in our training set. So as long as the true function $f^*(x) = \mathbb{E}(Y|X = x)$ is not very smooth, the nearest neighbors will not be informative for prediction.

Decomposing prediction error

A critical perspective for our course is that of dividing the prediction error / EPE of a predictive modeling approach into its components, which we will call bias and variance for squared error loss, and more generally approximation and estimation error for the general case.

We described predictive modeling as a process that inputs *random* training data T, runs it through a black-box modeling approach, outputs a model $\hat{f}$. We then get a new prediction point X_0, and we are interested in the prediction error $L(Y_0, \hat{f}(X_0))$. Considering all randomness, the quantity of interest is:

$$EPE(\text{modeling approach}) = \mathbb{E}_{T,X_0,Y_0} L(Y_0, \hat{f}(X_0)).$$

We can start by assuming $X_0 = x_0$ is fixed for simplicity, $L(u, v) = (u - v)^2$ squared error loss. Then we can write:

$$\begin{aligned}
\mathbb{E}_{T,Y_0}(Y_0 - \hat{f}(x_0))^2 &= \mathbb{E}_{T,Y_0}\left(Y_0 - \mathbb{E}(Y_0) + \mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0)) + \mathbb{E}(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2 = \\
&= \mathbb{E}_{T,Y_0}\left(Y_0 - \mathbb{E}(Y_0)\right)^2 + \left(\mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0))\right)^2 + \mathbb{E}\left(\mathbb{E}(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2 + \\
&\quad + \text{three cross terms that equal zero}
\end{aligned}$$

The cross terms:

$$2\mathbb{E}\left[\left(Y_0 - \mathbb{E}(Y_0)\right)\left(\mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0))\right)\right] \overset{\text{not random}}{=} 2\left(\mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0))\right)\mathbb{E}\left(Y_0 - \mathbb{E}(Y_0)\right) = 0$$

$$2\mathbb{E}_{T,Y_0}\left[\left(Y_0 - \mathbb{E}(Y_0)\right)\left(\mathbb{E}(\hat{f}(x_0) - \hat{f}(x_0))\right)\right] \overset{\text{independence}}{=} 2\mathbb{E}_{Y_0}\left(Y_0 - \mathbb{E}(Y_0)\right)\mathbb{E}_T\left(\mathbb{E}(\hat{f}(x_0) - \hat{f}(x_0))\right) = 0$$

$$2\mathbb{E}\left[\left(\mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0))\right)\left(\mathbb{E}(\hat{f}(x_0) - \hat{f}(x_0))\right)\right] \overset{\text{not random}}{=} 2\left(\mathbb{E}(Y_0) - \mathbb{E}(\hat{f}(x_0))\right)\mathbb{E}_T\left(\mathbb{E}(\hat{f}(x_0) - \hat{f}(x_0))\right) = 0$$

What we are left with:

$\sigma^2(x_0) = \mathbb{E}_{Y_0}\left(Y_0 - \mathbb{E}(Y_0)\right)^2$: *Irreducible* variance that does not depend on the modeling approach or T

$B(x_0) = \left(\mathbb{E}(Y_0) - \mathbb{E}_T(\hat{f}(x_0))\right)^2$: Squared *Bias* — how far is the mean prediction from the true mean?

$V(x_0) = \mathbb{E}_T\left(\mathbb{E}(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2$: Prediction *Variance* of the predicted value around its expectation

This is the famous **bias-variance decomposition** of squared prediction error. In addition to decomposing nicely from a mathematical perspective, it also gives us terms that clarify what are the sources of prediction error:

- Irreducible error that cannot be avoided due to data randomness

- Bias or *approximation error* that is due to the limited flexibility of our modeling approach (that cannot properly express $\mathbb{E}(Y_0)$ even on average)
- Variance or *estimation error* that measures the sensitivity/variability of the modeling approach due to the randomness in T

Important notes:

- We are usually not interested in the loss at specific x_0 but integrated over X: $EPE = E_X(\sigma^2(X)) + E_X(B(X)) + E_X(V(X))$.
- When using other loss functions than squared loss for prediction, we do not have this elegant mathematical decomposition, but the notions of approximation error and estimation error as underlying the performance is still important and relevant

Examples:

- For k-NN: having a small number of neighbors (and therefore small neighborhood) gives high variance, low bias vs. high number of neighbors (including far away) with less variance, higher bias. As the dimension p increases, both bias and variance can diverge.
- For linear regression: the variance increases roughly linearly with the number of variables p, while the bias decreases with p, as we will show