Statistical Learning, Fall 2024-25

Class notes 3

Prediction error decomposition

Bias-variance decomposition of squared prediction error for the model $Y = \mathbb{E}(Y|X) + \epsilon$, $\epsilon \sim (0, \sigma^2)$ (independent of X) :

$$\mathbb{E}\left(Y - \hat{f}(X)\right)^2 - \sigma^2 = \mathbb{E}_X\left(\mathbb{E}(Y|X) - \mathbb{E}(\hat{f}(X)|X)\right)^2 + \mathbb{E}_X \text{Var}(\hat{f}(X)|X) = \mathbb{E}_X(B(X)) + \mathbb{E}_X(V(X)).$$

We can now apply this to our "model" problems of least squares regression and k-NN. To start from a simple case, assume for least squares regression that the linear model is true $\mathbb{E}(Y|X) = X^T\beta$. For least squares we know $\hat{\beta} = \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\mathbb{Y}$, we can therefore write:

$$\hat{f}(x_0) = x_0^T\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\mathbb{Y} = x_0^T\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T(\mathbb{X}\beta + \epsilon) = x_0^T\beta + x_0^T\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\epsilon,$$

which immediately tells us that $\mathbb{E}(\hat{f}(X)|X = x_0) = x_0^T\beta$ and therefore $\mathbb{E}_X B(X) = 0$ and there is no bias. Note also that the first term is non-random given $X = x_0$.

To estimate the variance we use the law of total variation (we take a fixed $X = x_0$ to simplify notation):

$$\text{Var}(\hat{f}(x_0)) = \text{Var}(x_0^T\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\epsilon) = \text{Var}(A(\mathbb{X})\epsilon) = Var_{\mathbb{X}}\mathbb{E}\left(A(\mathbb{X})\epsilon \mid \mathbb{X}\right) + \mathbb{E}_{\mathbb{X}}\text{Var}\left(A(\mathbb{X})\epsilon \mid \mathbb{X}\right).$$

Because $\mathbb{E}(\epsilon) = 0$ and independence, we have for the first term:

$$\mathbb{E}\left(A(\mathbb{X})\epsilon \mid \mathbb{X}\right) \equiv 0 \Rightarrow \text{Var}_{\mathbb{X}}\mathbb{E}\left(A(\mathbb{X})\epsilon \mid \mathbb{X}\right) = 0.$$

Analyzing the second term, the inner expression gives:

$$V(x_0) = \mathbb{E}_{\mathbb{X}}\text{Var}\left(A(\mathbb{X})\epsilon \mid \mathbb{X}\right) = \mathbb{E}_{\mathbb{X}}A(\mathbb{X})Cov(\epsilon)A(\mathbb{X})^T = \sigma^2 A(\mathbb{X})A(\mathbb{X})^T = \sigma^2 x_0^T\mathbb{E}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}x_0.$$

Now integrating over the distribution of X we get:

$$\mathbb{E}_X(V(X)) = \sigma^2\mathbb{E}_{X,\mathbb{X}}X^T\left[\mathbb{X}^T\mathbb{X}\right]^{-1}X.$$

To calculate this expectation, note that by the law of large numbers: $\mathbb{X}^T\mathbb{X} = \sum_i X_i X_i^T \approx n\mathbb{E}XX^T$, and under mild conditions this also means $\left[\mathbb{X}^T\mathbb{X}\right]^{-1} \approx \frac{1}{n}\left[\mathbb{E}XX^T\right]^{-1}$. Using this we can write for large n:

$$\mathbb{E}_X(V(X)) \approx \frac{\sigma^2}{n}\mathbb{E}_X X^T\left[\mathbb{E}XX^T\right]^{-1}X \overset{(*)}{=} \sigma^2 Tr(\mathbb{E}_X XX^T\left[\mathbb{E}XX^T\right]^{-1}) \overset{(**)}{=} \frac{p\cdot\sigma^2}{n},$$

where $(*)$ uses the circular trace identity $Tr(ABC) = Tr(CAB)$ and $(**)$ uses the fact that we get the trace of an identity $p \times p$ matrix.

Putting everything together, we conclude that for true linear model, the prediction error of least squares regression is:

$$\mathbb{E}\left(Y - \hat{f}(X)\right)^2 \approx \sigma^2 + 0 + \frac{p \cdot \sigma^2}{n}.$$

Note the difference from the standard results in linear regression courses, where there is exact equality, and the variance formula does not depend on assuming linearity. The key for the difference is the *fixed-X* assumption taken in those courses, which we will get back to.

For k-NN, we can use the decomposition to derive various results, including the optimality result in the homework extra-credit problem, which we will get back to. Now we can derive a simpler but very interesting asymptotic result on 1-NN. Assume $f(X) = \mathbb{E}(Y|X)$ is "nicely behaved" (for example, obeys a Lipschitz condition $\|x_1 - x_2\| < \delta \Rightarrow |f(x_1) - f(x_2)| < C\delta$), and that X has a continuous distribution. For x_0 in the support of X denote by i_0 the index of its nearest neighbor in T, then from continuity of the distribution we have:

$$X_{i_0} \overset{\text{a.s.}}{\longrightarrow} x_0 \ , \quad \text{as } n \to \infty,$$

and therefore:

$$\mathbb{E}\hat{f}_{1-NN}(x_0) \overset{\text{a.s.}}{\longrightarrow} f(x_0) \ , \quad \text{meaning } B(x_0) \to 0,$$

with nice behavior this also guarantees $\mathbb{E}B(X) \overset{n\to\infty}{\longrightarrow} 0$, and only the variance component remains.

To quantify the variance component, we can again employ the law of total variation:

$$\text{Var}(\hat{f}(x_0)) = \text{Var}\mathbb{E}\left(\hat{f}(x_0) \mid \text{ location of } X_{i_0}\right) + \mathbb{E}\text{Var}\left(\hat{f}(x_0) \mid \text{ location of } X_{i_0}\right) = \text{Var}(f(X_{i_0})) + \sigma^2.$$

The second term is equal exactly to the variance of the noise σ^2 because we are conditioning on the location of the neighbor, hence it is simply the variance of Y at a given point. The first term converges to 0 because $X_{i_0} \to x_0$, so for large n, $f(X_{i_0})$ is approximately fixed at $f(x_0)$.

Putting all of this together we conclude that for $1 - NN$ and large enough training size n, we have:

$$\mathbb{E}\left(Y - \hat{f}(X)\right)^2 \approx 2\sigma^2,$$

twice the irreducible error (the minimal possible expected loss). Note that this is a very strong result and at first sight may seem to contradict the curse of dimensionality, which says that k-NN methods are problematic in high dimension. Of course, there is no contradiction due to the different asymptotics being employed ("big" p vs fixed p and $n \to \infty$).

Summarizing the results we proved here, with the result from the homework and the curse of dimensionality results, we can summarize our findings in the table below.

	p small, n big		p big, n not huge	
	Bias^2	Var	Bias^2	Var
Unbiased linear regresion	0	$\approx \frac{p}{n}\sigma^2$	0	$\approx \frac{p}{n}\sigma^2$
1-NN			COD	
	$\to 0$	$\to \sigma^2$	Big	Big
k-NN	With $k \to \infty, k/n \to 0$		COD	
	$\to 0$	$\to 0$	Big	May be controllable

Components of error in fixed-X linear regression

To get a more intuitive and geometric view of the topic, we can expose the dirty secret of regression courses: the fixed-X assumption. This assumption states that we are only interested in predicting at the same points $\mathbb{X}$ as in our training data (for new independent $\mathbb{Y}$ values of course). Furthermore, we assume that the values in $\mathbb{X}$ are given in advance and are not random. Writing this formally:

$$T = \mathbb{Y} = f(\mathbb{X}) + \epsilon \ , \quad \mathbb{Y}^{new} = f(\mathbb{X}) + \epsilon^{new} \ , \quad \text{Prediction error:} \frac{1}{n}\mathbb{E}_{\mathbb{Y},\mathbb{Y}^{new}} \|\mathbb{Y}^{new} - \hat{\mathbb{Y}}\|^2.$$

Note that now the "model" is simply a vector $\hat{\mathbb{Y}} \in \mathbb{R}^n$ of predictions for the specific points in $\mathbb{X}$.

This simplifies many aspects, for example we can calculate the prediction variance of least squares regression and find out that it is exactly $\sigma^2 p/n$, regardless if the linear model is true or not:

$$V = \frac{1}{n}\sum_{i=1}^{n} \text{Var}(\hat{Y}_i) = \frac{1}{n} tr(Cov(\mathbb{X}\hat{\beta})) = \frac{1}{n} tr(Cov(H\mathbb{Y})) = \frac{1}{n} tr(HCov(Y)H^T) = \frac{\sigma^2}{n} tr(HH^T) = \frac{\sigma^2}{n} tr(H) = \frac{p\sigma^2}{n},$$

where $H = \mathbb{X}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T$ is the hat matrix, and we use its symmetry ($H^T = H$), idemoptency ($HH = H$) and circular trace properties:

$$tr(HH) = tr(\mathbb{X}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\mathbb{X}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T) = tr(\mathbb{X}^T\mathbb{X}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}) = tr(I_p) = p.$$

The attached powerpoint presentation shows some mathematical derivations and especially geometric intuitions in this model.