MLISP: Machine Learning in Signal Processing

Lecture 6

Prof. V. I. Morgenshtern

Scribe: M. Solomon

Illustrations: The elements of statistical learning, Hastie, Tibshirani, Friedman

Agenda:

1. Bias and variance for linear regression
2. Summary: linear regression vs. nearest neighbors algorithm

1 Bias and variance for linear regression

Suppose now that the true relationship between $\mathbb{y}$ and $\mathbb{x}$ is linear and that there is noise

$$\mathbb{y} = \mathbb{x}^{\mathsf{T}}\boldsymbol{\theta} + \mathbb{n}$$

where $\mathbb{n} \sim \mathcal{N}(0, \sigma^2)$. In this lecture we will assume that the input variables $\mathbb{x}$ come from a random distribution, for example, they may be uniformly distributed in a box $[-1, 1]^p$ in p-dimensional space.

Let $\mathbf{x}_0$ be an arbitrary test point. Then the corresponding output is $\mathbb{y}_0 = \mathbf{x}_0^{\mathsf{T}}\boldsymbol{\theta} + \mathbb{n}_0$. Note that

$$\mathbb{E}[\mathbb{y}_0] = \mathbb{E}[\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\theta} + \mathbb{n}_0] = \mathbf{x}_0^{\mathsf{T}}\boldsymbol{\theta}.$$

Therefore, $\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\theta}$ is our ideal answer at test time. As before, our estimate at test time is

$$\hat{\mathbb{y}}_0 = \mathbf{x}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}$$

where

$$\hat{\boldsymbol{\theta}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\mathbb{y}.$$

Above, $\mathbb{X}$ is the input data matrix composed of rows $(\mathbb{x}^{(1)})^{\mathsf{T}}, \ldots, (\mathbb{x}^{(n)})^{\mathsf{T}}$, the data points in the training set; $\mathbb{y} = [\mathbb{y}^{(1)}, \ldots, \mathbb{y}^{(n)}]^{\mathsf{T}}$ are the output variables in the training set; n is the size of the training set. Clearly $\hat{\boldsymbol{\theta}}$ is random, even though we do not use our special font to denote that.

Now compute expected prediction error in this case:

$$\mathrm{EPE}(\mathbf{x}_0) = \mathbb{E}_{\mathbb{y}_0|\mathbb{x}_0=\mathbf{x}_0}\, \mathbb{E}_{T}(\mathbb{y}_0 - \hat{\mathbb{y}}_0)^2.$$

Above, $\mathbb{E}_T$ denotes expectation over the randomness in the training data generation, i.e. over noise and over the randomness in the data points $\mathbb{X}$. The expectation $\mathbb{E}_{\mathbb{y}_0|\mathbb{x}_0=\mathbf{x}_0}$ is with respect to noise at test time.

Let's simplify the expression for the expected prediction error. First add and subtract our ideal answer, $\mathbb{E}[\mathtt{y}_0] = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta}$:

$$\begin{aligned}
\mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathtt{y}_0 - \hat{\mathtt{y}}_0)^2 &= \mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathtt{y}_0 - \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} + \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \hat{\mathtt{y}}_0)^2 \\
&= \mathrm{Var}(\mathtt{y}_0|\mathtt{x}_0 = \mathbf{x}_0) + \mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \hat{\mathtt{y}}_0)^2 + 2\underbrace{\mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathtt{y}_0 - \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta})(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \hat{\mathtt{y}}_0)}_{0}\,.
\end{aligned}$$

To see that the last term is zero note that $\mathtt{y}_0 - \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} = \mathtt{n}_0$ is the test-time noise, which is zero mean and independent of $\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \hat{\mathtt{y}}_0$, which contains only the randomness of the training set.

Next, let's consider the second term:

$$\begin{aligned}
\mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \hat{\mathtt{y}}_0)^2 &= \mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\hat{\mathtt{y}}_0 - \mathbb{E}_T\,\hat{\mathtt{y}}_0 + \mathbb{E}_T\,\hat{\mathtt{y}}_0 - \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta})^2 \\
&= \mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\hat{\mathtt{y}}_0 - \mathbb{E}_T\,\hat{\mathtt{y}}_0)^2 + \mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathtt{y}}_0)^2 \\
&\quad + 2\underbrace{\mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\hat{\mathtt{y}}_0 - \mathbb{E}_T\,\hat{\mathtt{y}}_0)(\mathbb{E}_T\,\hat{\mathtt{y}}_0 - \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta})}_{0}\,. \qquad (1)
\end{aligned}$$

Above, $\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathtt{y}}_0$ is nonrandom and $\mathbb{E}_T(\hat{\mathtt{y}}_0 - \mathbb{E}_T\,\hat{\mathtt{y}}_0) = \mathbb{E}_T(\hat{\mathtt{y}}_0) - \mathbb{E}_T(\hat{\mathtt{y}}_0) = 0$; therefore the last term is zero.

Again, since $\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathtt{y}}_0$ is nonrandom, the second term in (1) can be written as:

$$\mathbb{E}_{\mathtt{y}_0|\mathtt{x}_0=\mathbf{x}_0}\, \mathbb{E}_T(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathtt{y}}_0)^2 = (\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathtt{y}}_0)^2 = (\mathrm{Bias}(\hat{\mathtt{y}}_0))^2\,.$$

This is the squared bias term. It measures how far $\mathbb{E}_T(\hat{\mathtt{y}}_0)$ is from the ideal estimate, $\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta}$. In this case, the squared bias is zero. This can be seen as follows. First observe that:

$$\hat{\mathtt{y}}_0 = \mathbf{x}_0^\mathsf{T}\hat{\boldsymbol{\theta}} = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} + \mathbf{x}_0^\mathsf{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

Now,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= (\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{y} - \boldsymbol{\theta} \\
&= (\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{X}\boldsymbol{\theta} + (\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{m} - \boldsymbol{\theta} \\
&= \boldsymbol{\theta} + (\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{m} - \boldsymbol{\theta} \\
&= (\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{m},
\end{aligned}$$

where $\mathtt{m}$ denotes the n-dimensional vector of noise in the training set. Therefore,

$$\mathbf{x}_0^\mathsf{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{x}_0^\mathsf{T}(\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\mathtt{m}$$

and so

$$\hat{\mathtt{y}}_0 = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} + \sum_{i=1}^{n} \mathtt{n}_i \left[\mathbf{x}_0^\mathsf{T}(\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\right]_i. \qquad (2)$$

Taking the expectation of the last expression and using that $\mathtt{n}_i$ is independent of the choice of training data points in $\left[\mathbf{x}_0^\mathsf{T}(\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\right]_i$, we conclude:

$$\mathbb{E}_T\,\hat{\mathtt{y}}_0 = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} + \sum_{i=1}^{n} \mathbb{E}_T[\mathtt{n}_i]\,\mathbb{E}_T\left[\mathbf{x}_0^\mathsf{T}(\mathtt{X}^\mathsf{T}\mathtt{X})^{-1}\mathtt{X}^\mathsf{T}\right]_i = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta}$$

so that

$$(\mathrm{Bias}(\hat{\mathbb{y}}_0))^2 = (\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta} - \mathbb{E}_T\,\hat{\mathbb{y}}_0)^2 = 0 \tag{3}$$

as claimed above.

Putting pieces together:

$$\mathrm{EPE}(\mathbf{x}_0) = \underbrace{\mathrm{Var}(\mathbb{y}_0|\mathbb{x}_0 = \mathbf{x}_0)}_{\mathrm{Var}(\mathbf{x}_0^\mathsf{T}\boldsymbol{\theta}+\mathbb{n}_0)=\sigma^2} + \mathbb{E}_{\mathbb{y}_0|\mathbb{x}_0=\mathbf{x}_0}\underbrace{\mathbb{E}_T(\hat{\mathbb{y}}_0 - \mathbb{E}_T(\hat{\mathbb{y}}_0))^2}_{\mathrm{Var}_T(\hat{\mathbb{y}}_0)} + \underbrace{(\mathrm{Bias}(\hat{\mathbb{y}}_0))^2}_{0}.$$

The first term is the additional variance σ^2 in the expected prediction error since our target is not deterministic. Let us analyze the second term, which is the variance in our predictions due to the randomness in the training set. Starting from (2) and using $\mathbb{E}_T(\hat{\mathbb{y}}_0) = \mathbf{x}_0^\mathsf{T}\boldsymbol{\theta}$, we have:

$$\begin{aligned}
\mathrm{Var}_T(\hat{\mathbb{y}}_0) &= \mathbb{E}_T\left(\sum_{i=1}^{n}\mathbb{n}_i\left[\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\right]_i\right)^2 \\
&= \sum_{i,j}\underbrace{\mathbb{E}_T[\mathbb{n}_i\mathbb{n}_j]}_{0 \text{ if } i\neq j}\,\mathbb{E}_T\left[\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\right]_i\left[\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\right]_j \\
&= \sigma^2\,\mathbb{E}_T\|\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\|_2^2 \\
&= \sigma^2\,\mathbb{E}_T\,\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbf{x}_0 \\
&= \sigma^2\,\mathbb{E}_T\,\mathbf{x}_0^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbf{x}_0.
\end{aligned}$$

We observe that the variance depends on the point $\mathbf{x}_0$.

To analyze this expression consider the case where we have a lot of training data, i.e. n is large and for simplicity assume $\mathbb{E}[\mathbb{x}] = \mathbf{0}$.

Under these assumptions,

$$\mathbb{X}^\mathsf{T}\mathbb{X} \to n\mathrm{Cov}(\mathbb{x}).$$

To see this consider components of this matrix.

For example,

$$\frac{1}{n}[\mathbb{X}^\mathsf{T}\mathbb{X}]_{1,1} = \frac{1}{n}\sum_i \mathbb{X}_{i1}\mathbb{X}_{i1} \to \mathbb{E}[\mathbb{x}_1^2],\ n\to\infty. \tag{4}$$

Above, $\mathbb{x}_i$ denotes the ith component of the vector $\mathbb{x}$ and the last step follows from independence of $\mathbb{X}_{1i}$ over i, and the law of large numbers.

Similarly,

$$\frac{1}{n}[\mathbb{X}^\mathsf{T}\mathbb{X}]_{1,2} = \frac{1}{n}\sum_i \mathbb{X}_{i1}\mathbb{X}_{i2} \to \mathbb{E}[\mathbb{x}_1\mathbb{x}_2],\ n\to\infty. \tag{5}$$

Next, assume that the test point $\mathbb{x}_0$ is drawn randomly from the same distribution as the points in the test set. Then we can average the expected prediction error over this choice:

$$\begin{aligned}
\mathbb{E}_{\mathbb{x}_0} \mathrm{EPE}(\mathbb{x}_0) &\to \sigma^2 + \mathbb{E}_{\mathbb{x}_0} \mathbb{x}_0^{\mathsf{T}} \mathrm{Cov}(\mathbb{x})^{-1} \mathbb{x}_0 \frac{\sigma^2}{n} \\
&= \sigma^2 + \frac{\sigma^2}{n} \mathbb{E}_{\mathbb{x}_0} \mathrm{tr}(\mathbb{x}_0^{\mathsf{T}} \mathrm{Cov}(\mathbb{x})^{-1} \mathbb{x}_0) \\
&= \sigma^2 + \frac{\sigma^2}{n} \mathbb{E}_{\mathbb{x}_0} \mathrm{tr}(\mathrm{Cov}(\mathbb{x})^{-1} \mathbb{x}_0 \mathbb{x}_0^{\mathsf{T}}) \\
&= \sigma^2 + \frac{\sigma^2}{n} \mathrm{tr}(\mathrm{Cov}(\mathbb{x})^{-1} \mathbb{E}_{\mathbb{x}_0} [\mathbb{x}_0 \mathbb{x}_0^{\mathsf{T}}]) \\
&= \sigma^2 + \frac{\sigma^2}{n} \mathrm{tr}(\underbrace{\mathrm{Cov}(\mathbb{x})^{-1} \mathrm{Cov}(\mathbb{x})}_{\mathbf{I}_p}) \\
&= \sigma^2 \frac{p}{n} + \sigma^2
\end{aligned}$$

where we used that $\mathrm{tr}(\mathbf{I}_p) = p$.

We see that the expected prediction error increases linearly as a function of p with the slope σ^2/n. Hence the expected prediction error is small if n is large or σ^2 is small. We have avoided the curse of dimensionality by putting heavy restrictions on the model class. We have no bias at all. However, if the model is wrong all bets are off.

2 Summary: linear regression vs. nearest neighbors algorithm

We conclude that linear regression has very desirable properties. If the true underlying model is linear, the linear regression has no bias and its variance increases only mildly with the number of variables in the model, p. However, if the underlying model is nonlinear, we might generate very large estimation errors because of the model mismatch. The nearest neighbors algorithm can flexibly adapt to any model. However, its bias is huge even in moderately small dimensions, $p = 10$. The art of machine learning is to design algorithms between these two extremes, well adapted to the problem at hand.