

Statistical Decision Theory, Least Squares, and Bias Variance Tradeoff

October 17, 2006

1 Supervised Learning Paradigm

Let x^i denote the *input* and y^i denote the *output*, which is what we trying to predict using x^i . The inputs are also known as the *covariates, features, predictors, or independent variables*. The outputs are also known as the *targets, response variables, or dependent variables*. A pair (x^i, y^i) is known as a training example and the data set used for learning, $T = \{(x^i, y^i) | i = 1, \dots, n\}$, is referred to as the training set. The superscript i indexes the example i .

Let \mathcal{X} be the space of input values and \mathcal{Y} be the space of output values. Typically \mathcal{Y} will be a subset of \mathcal{R} .

The goal in supervised learning is, given a *training set*, to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$. We refer to f as the *hypothesis* or predictive rule. We sometimes use f_T to denote hypothesis learned by the training algorithm given training example T .

2 Some Basics of Decision Theory

Let us assume we have some joint distribution over X, Y . Say we are interested in finding a function that minimizes the *least squares error*:

$$\mathbb{E}_{x,y}[(y - f(x))^2]$$

where x and y are random variables.

Remark 2.1: The optimal hypothesis f^* which minimizes the above error is the conditional expectation, namely

$$f^*(x) = E_y[y|x]$$

To see this note the following, let us examine the error for a specific choice of x . For this x , define \bar{y} to be $E_y[y|x]$,

$$\begin{aligned}
\mathbb{E}_y[(y - f(x))^2|x] &= \mathbb{E}_y[(y - \bar{y} + \bar{y} - f(x))^2|x] \\
&= \mathbb{E}_y[(y - \bar{y})^2|x] + \mathbb{E}_y[(\bar{y} - f(x))^2|x] \\
&\quad + 2\mathbb{E}_y[(y - \bar{y})(\bar{y} - f(x))|x] \\
&= \mathbb{E}_y[(y - \bar{y})^2|x] + \mathbb{E}_y[(\bar{y} - f(x))^2|x] \\
&\quad + 2(\bar{y} - f(x))\mathbb{E}_y[(y - \bar{y})|x] \\
&= \mathbb{E}_y[(y - \bar{y})^2|x] + \mathbb{E}_y[(\bar{y} - f(x))^2|x]
\end{aligned}$$

Where the last step follow from the definition of \bar{y} . Note that f does not effect the first term. Hence, the best choice of f is that which minimizes the second term, which is at $f(x) = \bar{y} = E_y[y|x]$.

To complete the argument, note that:

$$\mathbb{E}_{x,y}[(y - f(x))^2] = \mathbb{E}_x[\mathbb{E}_y[(y - f(x))^2|x]]$$

and we have found the minima of the inside term.

Alternatively, we might be interested in absolute loss:

$$\mathbb{E}_{x,y}[|y - f(x)|]$$

For this loss, one can show that best predictor is the conditional median, i.e.

$$f(x) = \text{Median } x|y$$

One should keep in mind what is trying to be predicted under a certain loss function.

3 Bias - Variance Tradeoff

Let us consider some learning algorithm and its expected prediction error:

$$\mathbb{E}_{x,y,T}[(y - f_T(x))^2]$$

Here, f_T is the hypothesis returned by the algorithm on training set T . Note that this function is random, where the randomness comes from the randomness in the training set.

Let us define, the mean prediction of the algorithm at point x to be:

$$\bar{f}(x) = \mathbb{E}_T[f_T(x)]$$

We can now decompose the error, at a fixed x , as follows:

$$\begin{aligned}
& \mathbb{E}_{y,T}[(y - f_T(x))^2] \\
= & \mathbb{E}_y[(y - \bar{y})^2] + \mathbb{E}_{y,T}[(\bar{f}(x) - f_T(x))^2] + \mathbb{E}_y[(\bar{y} - \bar{f}(x))^2] \\
= & \text{var}(y|x) + \text{var}_T(f(x)) + \text{bias}(f_T(x))
\end{aligned}$$

Let us interpret these terms. The first term, $\text{var}(y|x)$, is the output variance, which we have not control over. The second term, $\text{var}_T(f(x))$, is the variance of the hypothesis — determined by how the prediction varies around its average prediction. The final term is the bias squared, where the bias is difference between the average prediction and the true conditional mean. The first term is nonzero when the target is not deterministically related to x .

Hence, we have the following:

Remark 3.1: The expected prediction error is:

$$\begin{aligned}
& \mathbb{E}_{x,y,T}[(y - f_T(x))^2] \\
= & E_x[\text{var}(y|x) + \text{var}_T(f(x)) + \text{bias}(f_T(x))]
\end{aligned}$$

The proof of this is similar to the one above.

4 Examples

We will now examine two cases to understand issues about learning from a training set.

4.1 Nearest Neighbor

The *k-nearest neighbor rule* for prediction uses the k nearest points to predict the output. In the case of regression where $\mathcal{Y} = \mathcal{R}$ the rule says to just take the average of the k nearest points. For classification where $\mathcal{Y} = \{0, 1\}$, this rule says to take the majority vote of the nearest points.

Under mild regularity conditions, the error of the k nearest neighbor rule converges to the optimal error, as $n \rightarrow \infty$, $k \rightarrow \infty$ and $k/n \rightarrow 0$.

The major problem with the k nearest neighbor rule is that while its bias is very little (since it is averaging nearby points), the variance is very large.

4.2 Regression

A more constrained approach is to fit the data with a linear predictor of X . We will focus on this in the next lecture.