CSE 546: Machine Learning **Lecture 2**

Least Squares

Instructor: Sham Kakade

1 Supervised Learning and Regression

We observe data:

$$\mathcal{T} = (x_1, y_1), \ldots (x_n, y_n)$$

from some distribution. Our goal may be to predict the Y give some X. If Y is real, we may wish to learn the conditional expectation $\mathbb{E}[Y|X_i]$.

Typically, in supervised learning, we are interested in some notion of the our prediction loss. For example, in regression, the average squared error for a function f is:

$$L_{\text{squared error}}(f) = \mathbb{E}(f(X) - Y)^2$$

where the expectation is with respect to a random X, Y pair. (Note: sometimes the average error in machine learning is referred to as the risk.)

Note that minimizing the squared loss function also corresponds to doing maximum likelihood estimation under the model $\Pr(Y|X, f) = N(f(X), \sigma^2)$. To see this observe that,

$$-\log \Pr(Y|X, f) = -\log \sqrt{2\pi\sigma^2} + \frac{(f(X) - Y)^2}{2\sigma^2}$$

which is the square loss (up to a linear transformation).

Our goal is to use our training set $\mathcal{T}$ to estimate a function $\hat{f}$ which has low error. Also, note that the lowest possible squared error is achieved by:

$$f_*(X) = \mathbb{E}[Y|X]$$

which is the conditional expectation.

1.1 Risk (and some terminology clarifications)

A learning algorithm (or a decision rule) δ is a mapping from $\mathcal{T}$ to some hypothesis space. In this the case of regression it is a mapping from $\mathcal{T}$ to a function f. The notion of risk in statistics measures the quality of this procedure, on average.

Let f^* be the minimizer of L in some set $\mathcal{F}$, e.g.

$$f^* \in \arg\min_{f \in \mathcal{F}} L(f)$$

The *regret* of f (sometimes referred to as the loss) is defined as:

$$L(f) - L(f^*)$$

which is a measure of the sub-optimality of f.

The *risk* is some measure of the (average) performance of a decision rule; where, importantly, an expectation is take over the training set $\mathcal{T}$. One natural definition of the risk function is:

$$Risk(\delta) = \mathbb{E}_{\mathcal{T}}[L(\delta(\mathcal{T}))] - L(f_*)$$

Note that the expectation is over the training set. Other definitions may also be appropriate (though, technically, the risk should always refer to the performance of the decision rule δ).

2 Linear Regression

Suppose that $X \in \mathbb{R}^d$. Our prediction loss on our training set for a linear predictor is:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i \cdot w - Y_i)^2 = \frac{1}{n}\|\mathbf{X}w - \mathbf{Y}\|^2$$

where $\mathbf{X}$ (X in boldface) is defined to be the $n \times d$ matrix whose rows are X_i and $\mathbf{Y}$ (Y in boldface) is vector where $[Y_1, Y_2, \ldots Y_n]^\top$.

The least squares estimator using an outcome $\mathbf{Y}$ is just:

$$\hat{\beta} = \arg\min_w \frac{1}{n}\|\mathbf{Y} - \mathbf{X}w\|^2$$

The first derivative condition, often referred to as the *normal questions*, is that:

$$\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0$$

which is sometimes referred to as the *normal equations*.

The least squares estimator (the MLE) is then:

$$\hat{\beta}_{\text{least squares}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

3 Review: The SVD; the "Thin" SVD; and the pseudo-inverse

Theorem 3.1. *(SVD) Let $\mathbf{X} \in \mathbb{R}^{n\times d}$. there exists $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ orthogonal matrices (e.g. matrices with orthonormal rows and columns, so that $UU^\top = \mathrm{I}_n$ and $VV^\top = \mathrm{I}_d$ where I_k is the $k \times k$ identity matrix) such that:*

$$\mathbf{X} = \sum_i \lambda_i u_i v_i^\top = Udiag(\lambda_1, \ldots \lambda_{\min\{n,d\}})V^\top$$

where $diag(\cdot)$ is diagonal $\mathbb{R}^{n\times d}$ matrix and the λ_i's are referred to as the the singular values.

For $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} \in \mathbb{R}^n$, suppose that the equation:

$$\mathbf{X}\beta = \mathbf{Y}$$

has a unique solution. Then:

$$\beta = \mathbf{X}^{-1}\mathbf{Y}$$

where $\mathbf{X}^{-1}$ is the inverse of $\mathbf{X}$ (it exists and is unique since we have assume the linear system has a unique solution). In regression, there is typically noise, and we find a β which minimizes:

$$\|\mathbf{X}\beta - \mathbf{Y}\|^2$$

Clearly, if there is no noise, then a solution is given by $\beta = \mathbf{X}^{-1}\mathbf{Y}$, assuming no degeneracies. In general though, the minimizer of this error, referred to as the *least squares estimator*, is:

$$\beta = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}\mathbf{Y}\,. \tag{1}$$

Furthermore, Equation 1 above only holds if $\mathbf{X}$ is of rank d (else $\mathbf{X}^\top\mathbf{X}^{-1}$ would not be invertible).

Now let us define the Moore-Penrose pseudo-inverse.

First, let us define the 'thin' SVD.

Definition 3.2. *We say $\mathbf{X} = UDV^\top$ is the "thin" SVD of $\mathbf{X} \in \mathbb{R}^{n\times p}$ if: $U^{n\times r}$ and $V^{p\times r}$ have orthonormal columns (e.g. where r is the number of columns) and $D \in \mathbb{R}^{r\times r}$ is diagonal, with all it's diagonal entries being non-zero. Here, r is the* rank *of $\mathbf{X}$.*

Now we define the pseudo-inverse as follows:

Definition 3.3. *Let $\mathbf{X} = UDV^\top$ be the thin SVD of $\mathbf{X}$. The Moore-Penrose pseudo-inverse of $\mathbf{X}$, denoted by $\mathbf{X}^+$, is defined as:*

$$\mathbf{X}^+ = VD^{-1}U^\top$$

Let us make some observations:

1. First, if $\mathbf{X}$ is invertible (so $\mathbf{X}$ is square) then $\mathbf{X}^+ = \mathbf{X}^{-1}$.
2. Suppose that $\mathbf{X}$ isn't square and that $\mathbf{X}w = \mathbf{Y}$ has a (unique) solution, then $w = \mathbf{X}^+\mathbf{Y}$.
3. Now suppose that $\mathbf{X}w = \mathbf{Y}$ has (at least one) solution. Then one solution is given by $w = \mathbf{X}^+\mathbf{Y}$. This solution is the minimum norm solution w.
4. (geometric interpretation) The matrix $\mathbf{X}^+$ maps any point in the range of $\mathbf{X}$ to the minimum norm point in the domain.

Using this terminology, we can write the least squares estimator in a more interpretable way:

Lemma 3.4. *The least squares estimator is:*

$$\beta = \mathbf{X}^+\mathbf{Y}$$

(Note that the above is alway a minimizer, while the solution provided in Equation 1 only holds if $\mathbf{X}^\top\mathbf{X}$ is invertible, in which case the minimizer is unique).

4 Analysis: what is the risk?

We will return to this in the next lecture.

5 What about if $d > n$?

We will examine this in the next few lectures.