Least Squares

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) \( \delta \) 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 \( F \), e.g.
\[ f^* \in \arg \min_{f \in F} L(f) \]

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:

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

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 $\delta$).

## 2 Linear Regression

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

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

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

The least squares estimator using an outcome $Y$ is just:

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

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

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

which is sometimes referred to as the normal equations.

The least squares estimator (the MLE) is then:

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

## 3 Review: The SVD; the “Thin” SVD; and the pseudo-inverse

**Theorem 3.1. (SVD)** Let $X \in \mathbb{R}^{n \times d}$, 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 $UU^\top = I_n$ and $VV^\top = I_d$ where $I_k$ is the $k \times k$ identity matrix) such that:

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

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

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

$$X\beta = Y$$

has a unique solution. Then:

$$\beta = X^{-1}Y$$
where $X^{-1}$ is the inverse of $X$ (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 $\beta$ which minimizes:

$$||X\beta - Y||^2$$

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

$$\beta = (X^TX)^{-1}XY.$$ (1)

Furthermore, Equation 1 above only holds if $X$ is of rank $d$ (else $X^TX$ 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 $X = UDV^T$ is the “thin” SVD of $X \in \mathbb{R}^{n \times p}$ 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 $X$.

Now we define the pseudo-inverse as follows:

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

$$X^+ = VD^{-1}U^T$$

Let us make some observations:

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

$$\beta = X^+Y$$

(Note that the above is always a minimizer, while the solution provided in Equation 1 only holds if $X^TX$ is invertible, in which case the minimizer is unique).