CS5014: Machine Learning, Revision Notes

Gergely Flamich

May 10, 2018

1 Overview

In ML, we usually have some data in the following general setup:

- **Features**: literally any sort of observations form the real world, collected conveniently into a matrix:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

  where each column $\mathbf{x}^{(i)}$ contains m observations of the i-th feature

- **Labels**: In supervised learning we also have another set of data, called the ground truth of the object we are trying to predict from the features. These are conveniently colleted into a vector $\mathbf{y}$ with m elements, each y_i corresponding to a set of observations, row $\mathbf{x}_i$ in X.

The general assumption for supervised learning is that each feature is associated with a random variable X_i over some distributions and the ground truth is associated with a random variable Y over some distribution. It is usually assumed that we know the distributions of X_i and Y is hidden. fianlly, we assume that there is some function f such that

$$Y = f(\mathbf{X}, \theta) + \epsilon$$

for some model paramters θ and some small error term ϵ. Then, the way we will perform prediction is by using the distribution

$$\hat{Y} = f(\mathbf{X}, \theta).$$

f is usually referred to as the **model function**.

2 Regression Problems

One of the simplest models f relies on the further assumption that Y depends linearly on $\mathbf{X}$. In particular,

$$\hat{Y} = f(\mathbf{X}, \theta) = \mathbf{X}\theta + b$$

where b is called the **bias**. No Then, we can quantify the goodness-of-fit of the model, by the squared error, specified by

$$L(\theta, b) = \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right|^2$$

Then, to find the optimal set of parameters theta, we can differentiate L:

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \frac{\partial}{\partial \theta} \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right|^2 \\
&= 2 \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right|^2 \frac{\partial}{\partial \theta} \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right| \\
&= 2 \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right| \frac{(\mathbf{y} - f(X, \theta, b))^T}{2 \left|\left|\mathbf{y} - f(X, \theta, b)\right|\right|} \frac{\partial}{\partial \theta} \mathbf{y} - f(X, \theta, b) \\
&= (\mathbf{y} - X\theta)^T \cdot -X
\end{aligned}$$

Now to find the minimum, set $\frac{\partial L}{\partial \theta} = 0$.

$$\begin{aligned}
0 &= -\mathbf{y}^T X + \theta^T X^T X \\
\mathbf{y}^T X &= \theta^T X^T X \\
\mathbf{y}^T X (X^T X)^{-1} &= \theta^T
\end{aligned}$$

Hence

$$\theta = (X^T X)^{-1} X^T \mathbf{y}$$

The above is called the **normal equation**. An alternative way to solve the above problem is **gradient descent**. Iterate

$$\theta \leftarrow \theta - \alpha \cdot \nabla L$$

for some appropriate $\alpha \in \mathbb{R}$. Then it can be shown that the θ converges to the minimal solution. α is called the **learning rate**.

3 SVMs

4 Neural Networks and Deep Learning