Lecture3 Note

Zongcheng Chu

January 2019

1 Recall from last lecture

Expected Risk:

$$R(h) = E_{(x,y)\sim Pr} L(h(x), y)$$
$$= E_{(x,y)\sim Pr} \mathbf{1}[h(x) \neq y]$$

Bayes optimal classifier:

$X = (x_1, x_2, x_3...)$ and $y \in (c_1, c_2, c_3...)$, if we know the probability of $Pr(y|x)$, then the optimal classifier would be:

$$f^*(x) = \arg\min_{c\in[C]} Pr(c|x)$$

And also:

$$R(f^*) \leq R(f), \forall f$$

2 New materials today

Determinstic Case:

(x1,x2)	y	$Pr(y = 0\|x)$	$Pr(y = 1\|x)$
(0,0)	1	0.3	0.7
(0,1)	0	0.6	0.4
(1,0)	0	0.8	0.2
(1,1)	1	0.1	0.9

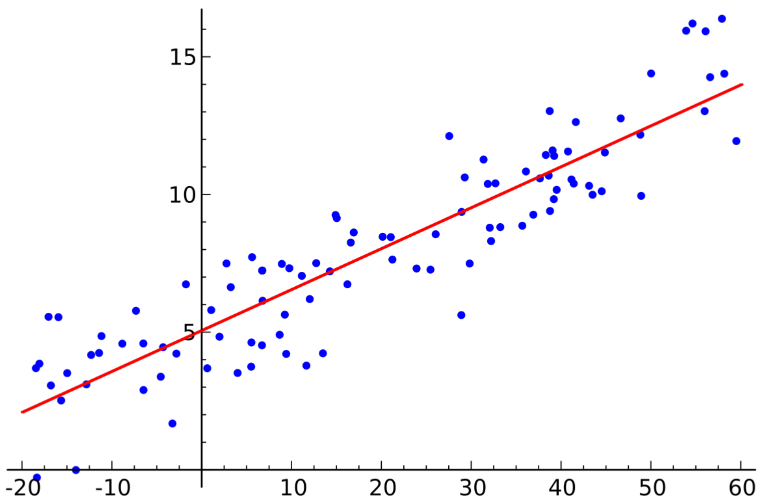
In this case, if we are given $X = (0, 0)$, we are trying to find out the most likely category corresponding to the given X. By looking up the table, y=1 has the greatest probability which is 0.7. This is how Bayes optimal classifier works.

Theorem(Cover Hart 1967)
f_N is a 1-NN binary classifier trained using N data points. Then we have:

$$R(f^*) \leq \lim_{N\to\infty} E[R(f_N)] \leq 2\mathrm{R}(\mathrm{f}^*)$$

But it is only in theory. In practice, (1) N can not be ∞. (2) Not all models are determinstic.

Linear regression



$$
\begin{aligned}
\mathbf{Y} &= \mathbf{WX} + \mathbf{b} \\
&= \mathrm{w}_1 \ * \ x1 \ + \ w_2 \ * \ x_2 \ + \ \mathbf{b} \\
&= [\mathrm{w}_1, w_2]*[x,x_2]^T+[b_1, b_2]^T \\
&= [\mathrm{w}_1, w_2, b]*[x_1, x_2, 1]^T
\end{aligned}
$$

Our training samples are: $D^{train} = \{(x_1, y_1), (x_2, y_2)...\}$,here linear regression follows the pattern:

$$y_i = W^T X_i + \Sigma_i$$

Where $\Sigma_i \sim N(0, \sigma^2)$ and $y_i \sim N(W^T X_i, \sigma^2)$

We find the optimal W and b by convert the problem into a Least Square problem. Why? Because we are tring to maximize the likelyhood function:

$$
\begin{aligned}
L &= \Pi_{i=1}^{N} Pr(y_i|x_i) \\
&= \Pi_{i=1}^{N} (\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(w^T x_i - y_i)^2}{2\sigma^2})) \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^2} exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (w^T x_i - y_i)^2)
\end{aligned}
$$

$(W^TX_i - y_i)^2$ is the square loss here, and our target is to minimize it to increase the likelyhood function. In practice, we have two different ways to find the optimal W^* and b^*.

Gradient Descent

The function we try to minimize is:

$$RSS = (W^TX_i - y_i)^2$$

$\nabla RSS = 2\sum_i X_i(X_iW - y_i) \propto (\sum_i(X_iX_i^T)W - \sum_i X_iy_i$

$$\text{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ . \\ . \\ . \\ X_n^T \end{bmatrix} \quad \text{Y} = \begin{bmatrix} y_1^T \\ y_2^T \\ . \\ . \\ . \\ y_n^T \end{bmatrix}$$

So,

$$\text{RSS} = (XX^T)W - X^TY$$

$$\text{W}^* = (\text{X}^TX)^{-1}X^TY$$

Matrix Deduction

$$\begin{aligned}
RSS(W) &= \sum_i (W^TX_i - y_i)^2 \\
&= ||XW - Y||^2 \\
&= (XW - Y)^T(XW - Y) \\
&= W^T(X^TX)W - Y^TXW - W^TX^TY + Y^TY \\
&= W^T(X^TX)W - Y^TXW - W^T(X^TX)(X^TX)^{-1}X^TY + Y^TX(X^TX)^{-1}X^TY \\
&= (W^T(X^TX) - Y^TX)W - (W^T(X^TX) - Y^TX)(X^TX)^{-1}X^TY \\
&= (W^T(X^TX) - Y^TX)(W - (X^TX)^{-1}X^TY) \\
&let \quad Y^TX = Y^TX(X^TX)^{-1}(X^TX) \\
&= (W - (X^TX)^{-1}X^TY)(X^TX)(W - (X^TX)^{-1}X^TY) \\
&= \mu^T(X^TX)\mu \\
&= ||x\mu||_2^2 \geq 0
\end{aligned}$$

$||x\mu||_2^2 = 0$,iff μ=0

So, $W^* = (X^TX)^{-1}X^TY$