# CIS 520, Machine Learning, Fall 2011: Assignment 3

Xiaojun Feng

October 12, 2011

## 1 Linear regression and LOOCV

1. Since $\hat{w} = (X^TX)^{-1}X^TY$, the time complexity of computing w is $m^2(n-1)+m^3+m^2(n-1)+m(n-1)$; also $\hat{Y} = X\hat{w}$,hence the time complexity of computing y is $m^2$. $LOOCV = \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(-i)})^2$,hence time complexity of computing is 1; thus the complexity of computing LOOCV is $m^2(n-1) + m^3$

2. $\hat{Y}_i$ could be represented as $\hat{Y}_i = \sum_{j=1}^{n} H_{ij}Y_i$

3. Since we have

$$\begin{aligned} SSE_z &= \sum_{j=1}^{n}(Z_j - \hat{Z}_j)^2 \\ &= \sum_{j\neq i}(Z_j - \hat{Z}_j)^2 + \sum_{j=i}(Z_j - \hat{Z}_j)^2 \\ &= \sum_{j\neq i}(Y_j - \hat{Z}_j)^2 + \sum_{j=i}(\hat{Y}_i^{(-i)} - \hat{Z}_j)^2 \end{aligned}$$

if $\hat{Z}_j = \hat{Y}^{(-i)}$,then

$$\begin{aligned} SSE_z &= \sum_{j\neq i}(Y_j - \hat{Y}_j^{(-i)})^2 + \sum_{j=i}(\hat{Y}_i^{(-i)} - \hat{Y}_j^{(-i)})^2 \\ &= \sum_{j\neq i}(Y_j - \hat{Y}_j^{(-i)})^2 \end{aligned}$$

, and we know LOOCV minimize $\sum_{j\neq i}(Y_j - \hat{Y}_j^{(-i)})^2$, thus $\hat{Y}^{(-i)}$minimize SSE for Z.

4. $\hat{Y}_i^{(-i)}$ could be represented as $\hat{Y}_i^{(-i)} = \sum_{k=1}^{n} H_{jk}Z_j$

5. Since we have $\hat{Y}_i - \hat{Y}_i^{(-i)} = \sum_{j=1}^{n} H_{ij}Y_i - \sum_{k=1}^{n} H_{jk}Z_j$, regarding to the definition of $Z_j$,

$$\begin{aligned} \sum_{j=1}^{n} H_{ij}Y_i - \sum_{k=1}^{n} H_{jk}Z_j &= \sum_{j\neq i} H_{jk}Y_j - \sum_{j\neq i} H_{jk}Y_j + H_{ii}Y_i\text{-}H_{ii}\hat{Y}_i^{(-i)} \\ &= H_{ii}Y_i\text{-}H_{ii}\hat{Y}_i^{(-i)} \end{aligned}$$

6. Based on the calculation above $\hat{Y}_i - \hat{Y}_i^{(-i)} = H_{ii}Y_i\text{-}H_{ii}\hat{Y}_i^{(-i)}$, thus $\hat{Y}_i^{(-i)} = \frac{\hat{Y}_i - H_{ii}Y_i}{1\text{-}H_{ii}}$, and

$$\begin{aligned} LOOCV &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(-i)})^2 \\ &= \sum_{i=1}^{n}(\frac{Y_i - \hat{Y}_i}{1\text{-}H_{ii}})^2 \end{aligned}$$

## 2 Logistic regression and Naive Bayes

1. Objective function.
   (a) Naive Bayes (g)
   (b) Logistic regression (c)

2. Prove logistic regression from Naive Bayes$P(Y=1|X) = \frac{1}{1+\exp\{w_0+w^TX\}}$

$$
\begin{aligned}
P(Y=1|X) &= \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)} \\
&= \frac{1}{1+\exp(\log\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}
\end{aligned}
$$

Now let's plug in our definitions:

$$
\begin{aligned}
P(Y=1) &= \pi \\
P(X_j|Y=1) &= \theta_{j1}^{X_j}(1-\theta_{j1})^{1-X_j} \\
P(X_j|Y=0) &= \theta_{j0}^{X_j}(1-\theta_{j0})^{1-X_j}
\end{aligned}
$$

$$
\begin{aligned}
\log\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)} &= \log\frac{P(Y=0)}{P(Y=1)} + \sum_j \log\frac{P(X_j|Y=0)}{P(X_j|Y=1)} \\
&= \log\frac{1-\pi}{\pi} + \sum_j \log\frac{\theta_{j1}^{X_j}(1-\theta_{j1})^{1-X_j}}{\theta_{j0}^{X_j}(1-\theta_{j0})^{1-X_j}} \\
&= \log\frac{1-\pi}{\pi} + \sum_j \log\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}X_j + \sum_j \log\frac{1-\theta_{j0}}{1-\theta_{j1}} \\
&= w_0 + w^TX
\end{aligned}
$$

## 3 Double-counting the evidence

1. 5 parameters will be needed, P(Y=T), P($X_1 = T|Y = T$), $P(X_1 = T|Y = F)$, $P(X_2 = T|Y = F)$, $P(X_2 = T|Y = T)$.

2.

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| T | T | T |
| T | F | T |
| F | T | T |
| F | F | F |

3. Naive Bayes error rate
   (a) since$P(X_1=T,X_2=T,Y=F) = P(X_1=T|Y=F)P(X_2=T|Y=F)P(Y=F) = 0.015$
   $P(X_1=T,X_2=F,Y=F) = P(X_1=T|Y=F)P(X_2=F|Y=F)P(Y=F) = 0.135$
   $P(X_1=F,X_2=T,Y=F) = P(X_1=F|Y=F)P(X_2=T|Y=F)P(Y=F) = 0.035$
   $P(X_1=F,X_2=F,Y=T) = P(X_1=F|Y=T)P(X_2=F|Y=T)P(Y=T) = 0.05.$

$$
\begin{aligned}
error\ rate = &\sum_{X_1,X_2,Y} 1(Y \neq f(X_1,X_2))P(X_1,X_2,Y) \\
&= P(X_1=T,X_2=T,Y=F) + P(X_1=T,X_2=F,Y=F) \\
&\quad + P(X_1=F,X_2=T,Y=F) + P(X_1=F,X_2=F,Y=T) \\
&= 0.235
\end{aligned}
$$

So, the error rate is 0.235.

(b) $P(X_1 = T, Y = F) = P(X_1 = T|Y = F)P(Y = F) = 0.15$
$P(X_1 = F, Y = T) = P(X_1 = F|Y = T)P(Y = T) = 0.1$
so, error rate is 0.25.

(c) $P(X_2 = T, Y = F) = P(X_2 = T|Y = F)P(Y = F) = 0.05$
$P(X_2 = F, Y = T) = P(X_2 = F|Y = T)P(Y = T) = 0.25$
so, error rate is 0.30.

(d) The error rate is lower using $X_1, X_2$together, because we use more independent features to classify Y, which makes the classification more accurately.

4. KL divergence for Multinomials.

(a) No, $X_2$and $X_2$ are not conditionally independent given Y, since $P(X_2, X_3|Y) = P(X_2|Y)\ and\ P(X_2|Y) = P(X_3|Y)$, so apparently, $P(X_2, X_3|Y) \neq P(X_2|Y)P(X_3|Y)$.

(b) After adding the $X_2$, the classification result becomes

| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| T | T | T | T |
| T | F | F | F |
| F | T | T | T |
| F | F | F | F |

$P(X_1 = T, X_2 = T, X_3 = T, Y = F) = P(X_1 = T|Y = F)P(X_2 = T, X = T|Y = F)P(Y = F) = 0.015$
$P(X_1 = T, X_2 = F, X_3 = F, Y = T) = P(X_1 = T|Y = T)P(X_2 = F, X_2 = F|Y = T)P(Y = F) = 0.2$
$P(X_1 = F, X_2 = T, X_3 = T, Y = F) = P(X_1 = F|Y = F)P(X_2 = T, X_3 = T|Y = F)P(Y = F) = 0.035$
$P(X_1 = F, X_2 = F, X_2 = F, Y = T) = P(X_1 = F|Y = T)P(X_2 = F, X_3 = F|Y = T)P(Y = T) = 0.05$
so, error rate is 0.3.

5. This is because, $X_2$and $X_3$are actually not independent, but when using Naive Bayes to make decision, they are treated as independent.

6. No, logistic regression doesn't suffer from the same problem, since it calculate p(y|x) directly by estimating the weight.

7. Extra credit

(a) Decision rule

$$\frac{P(Y = T|X_1 = T, X_2 = F, X_3 = F)}{P(Y = F|X_1 = T, X_2 = F, X_3 = F)} > 1$$

$$\frac{\frac{P(X_! = T|Y = T)P(X_2 = F|Y = T)P(X_3 = F|Y = T)P(Y = T)}{P(X_1)P(X_2)P(X_3)}}{\frac{P(X_! = T|Y = F)P(X_2 = F|Y = F)P(X_3 = F|Y = T)P(Y = F)}{P(X_1)P(X_2)P(X_3)}} > 1$$

$$\frac{\frac{P(Y = T|X_1 = T)P(X_1 = T)P(Y = T|X_2 = F)P(X_2 = F)P(Y = T|X_3 = F)P(X_3 = F)P(Y = T)}{P(Y = T)P(Y = T)P(Y = T)}}{\frac{P(Y = T|X_1 = T)P(X_1 = T)P(Y = T|X_2 = F)P(X_2 = F)P(Y = T|X_2 = F)P(X_2 = F)P(Y = F)}{P(Y = T)P(Y = T)}} > 1$$

$$\frac{P(Y=T|X_1=T)P(Y=T|X_2=F)P(Y=T|X_3=F)P(Y=T)}{P(Y=F|X_1=T)P(Y=F|X_2=F)P(Y=F|X_3=F)P(Y=F)} > 1$$

$$\frac{pq^2}{(1-p)(1-q)^2} > 1$$

$$pq^2 > (1-q)^2 - p(1-q)^2$$

$$p > \frac{(1-q)^2}{q^2+(1-q)^2}$$

(b) Real decision rule

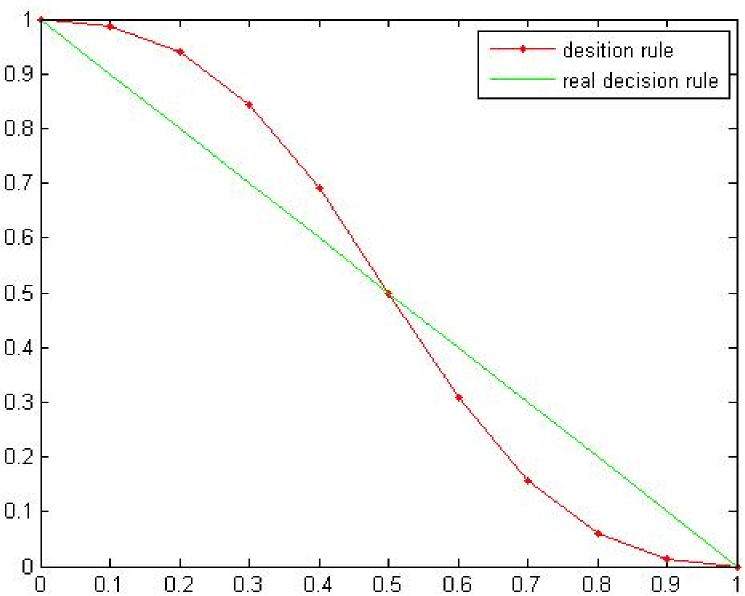
$$\frac{P(Y=T|X_1=T,X_2=F,X_3=F)}{P(Y=F|X_1=T,X_2=F,X_3=F)} > 1$$

$$\frac{P(Y=T|X_1=T)P(Y=T|X_2=F)P(Y=T)}{P(Y=F|X_1=T)P(Y=F|X_2=F))P(Y=F)} > 1$$

$$\frac{pq}{(1-p)(1-q)} > 1$$

$$p > 1-q$$

(c) figure




# 4 Boosting

## 4.1 Analyzing the training error of boosting

1. Show, $\frac{1}{m}\sum_{i=1}^{m} I(H(x_i) \neq y_i) \leq \frac{1}{m}\sum_{i=1}^{m} \exp(-f(x_i)y_i)$
   If $H(x_i) = y_i,\ I(H(x_i) \neq y_i) = 0 \leq \exp(-f(x_i)y_i)$,
   if $H(x_i) \neq y_{i,}\ -f(x_i)y_i > 0,\ I(H(x_i \neq y_i)) = 1 \leq \exp(-f(x_i)y_i)$
   so, $\frac{1}{m}\sum_{i=1}^{m} I(H(x_i) \neq y_i) \leq \frac{1}{m}\sum_{i=1}^{m} \exp(-f(x_i)y_i)$.