CIS 520, Machine Learning, Fall 2018: Assignment 3

Wentao He

October 2, 2018

Collaborators: N/A

1 Naïve Bayes as a Linear Classifier

1. Based on the question, since we are assuming that all the attributes of each instance $\mathbf{x}$ are conditionally independent given y, we know that:$\mathbf{Pr}(\mathbf{x}|y=1) = \prod_{i=1}^{n} \mathbf{Pr}(x_i|y=1)$, so that

$$\mathbf{Pr}(x_i|y=1) = \begin{cases} \alpha_i & \text{if } x_i = 1 \\ (1-\alpha_i) & \text{if } x_i = 0. \end{cases}$$

Therefore we know that the conditional probability of $\mathbf{x}$ given y can be written as $\mathbf{Pr}(\mathbf{x}|y=1) = \prod_{i=1}^{n} \alpha_i^{x_i} \cdot (1-\alpha_i)^{(1-x_i)}$, since $\alpha_i^{x_i} \cdot (1-\alpha_i)^{(1-x_i)}$ is just another way to write the aforementioned equations in the bracket. The same set of equations can be written for $y = -1$. Therefore the conditional probability of $\mathbf{x}$ given y can also be written as $\mathbf{Pr}(\mathbf{x}|y=-1) = \prod_{i=1}^{n} \beta_i^{x_i} \cdot (1-\beta_i)^{(1-x_i)}$.

2. **The maximum likelihood estimates (MLE) of** p:

Given data $D = \{(\mathbf{x_1}, y_1), \cdots, (\mathbf{x_m}, y_m)\}$, the log-likelihood function L of p can be written as $L(D;p) = C_1 + \sum_{i=1}^{m} \frac{1+y_i}{2} \cdot \log(p) + \frac{1-y_i}{2} \cdot \log(1-p)$, and the term C_1 only depends on α's and β's. From the above equation, the MLE of $\hat{p}$ can be written as $\hat{p} = \text{argmax}_p\, L(D;p)$. The derivative of $L(D;p)$ with respect to p can be written as $\frac{dL(D;p)}{dp} = \sum_{i=1}^{m} \frac{1+y_i}{2p} - \frac{1-y_i}{2(1-p)}$. When this derivative equals 0,

$$\boxed{\hat{p} = \frac{1}{m}\sum_{i=1}^{m} \frac{1+y_i}{2} = \frac{\text{number of positive data points in D}}{m}}.$$

The maximum likelihood estimates (MLE) of α_i:

Given data $D = \{(\mathbf{x_1}, y_1), \cdots, (\mathbf{x_m}, y_m)\}$, the log-likelihood function L of α_i can be written as $L(D;\alpha_i) = C_2 + \sum_{j:y_j=1} x_{ji} \cdot \log(a_i) + (1-x_{ji}) \cdot \log(1-\alpha_i)$, and the term C_2 only depends on p, β's, and $\alpha_{i'}$ when $i' \neq i$. The derivative of $L(D;\alpha_i)$ with respect to α_i can be written as $\frac{dL(D;\alpha_i)}{d\alpha_i} = \sum_{j:y_j=1} \frac{x_{ji}}{\alpha_i} - \frac{1-x_{ji}}{1-\alpha_i}$. When this derivative equals 0,

$$\boxed{\hat{\alpha_i} = \frac{\sum_{j:y_j=1} x_{ji}}{\sum_{j:y_j=1} 1} = \frac{\text{number of positive data points in D where the i-th component equals to 1}}{\text{number of positive data points in D}}}.$$

The maximum likelihood estimates (MLE) of β_i:

Similar to the above derivation of the maximum likelihood estimates (MLE) of α_i,

$$\boxed{\hat{\beta}_i = \frac{\sum\limits_{j:y_j=-1} x_{ji}}{\sum\limits_{j:y_j=-1} 1} = \frac{\text{number of negative data points in D where the } \textbf{i}\text{-th component equals to 1}}{\text{number of negative data points in D}}}.$$

3. Using the proposed equation, we know that

$$h(\mathbf{x}) = \begin{cases} 1 & \text{when } \hat{\mathbf{Pr}}(1|x) > \hat{\mathbf{Pr}}(-1|x) \\ -1 & \text{when otherwise.} \end{cases}$$

so that

$$h(\mathbf{x}) = \underset{y\in\{\pm 1\}}{\operatorname{argmax}} \hat{\mathbf{Pr}}(y|\mathbf{x}) = \begin{cases} 1 & \text{when } \hat{\mathbf{Pr}}(1|x) > \hat{\mathbf{Pr}}(-1|x) \\ -1 & \text{when otherwise.} \end{cases}$$

4. Using Bayes rule, we get

$$\begin{aligned} h(x) &= \operatorname{sign}(\frac{\hat{\mathbf{Pr}}(1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|1)}{\hat{\mathbf{Pr}}(\mathbf{x})} - \frac{\hat{\mathbf{Pr}}(-1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|-1)}{\hat{\mathbf{Pr}}(\mathbf{x})}) \\ &= \hat{\mathbf{Pr}}(1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|1) - \hat{\mathbf{Pr}}(-1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|-1) \\ &= \operatorname{sign}(\log(\hat{\mathbf{Pr}}(1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|1)) - \log(\hat{\mathbf{Pr}}(-1)\cdot\hat{\mathbf{Pr}}(\mathbf{x}|-1))) \\ &= \operatorname{sign}(\log(\hat{p}) + \sum_{i=1}^{n}(x_i\cdot\log(\hat{\alpha}_i) + (1-x_i)\cdot\log(1-\hat{\alpha}_i)) - \\ &\quad \log(1-\hat{p}) - \sum_{i=1}^{n}(x_i\cdot\log(\hat{\beta}_i) + (1-x_i)\cdot\log(1-\beta_i))) \\ &= \operatorname{sign}(\sum_{i=1}^{n} x_i \log(\frac{\hat{\alpha}_i\cdot(1-\hat{\beta}_i)}{\hat{\beta}_i\cdot(1-\hat{\alpha}_i)}) + \sum_{i=1}^{n}\log\frac{1-\hat{\alpha}_i}{1-\hat{\beta}_i} + \log\frac{\hat{p}}{1-\hat{p}}) \\ &= \operatorname{sign}(\mathbf{w}^\top\mathbf{x} + b), \end{aligned}$$

so that we get

$$\begin{aligned} w_i &= \boxed{\log\frac{\hat{\alpha}_i\cdot(1-\hat{\beta}_i)}{\hat{\beta}_i\cdot(1-\hat{\alpha}_i)}} \\ b &= \boxed{\log\frac{\hat{p}}{1-\hat{p}} + \sum_{i=1}^{n}\log\frac{1-\hat{\alpha}_i}{1-\hat{\beta}_i}}. \end{aligned}$$

2 Multiclass Logistic Regression

1. The likelihood function can be written as

$$
\begin{aligned}
l(\mathbf{w}_1, \cdots, \mathbf{w}_C) &= \prod_{m=1}^{M} \mathbf{P}(Y_m = y_m | \mathbf{X}, \mathbf{w}) \\
&= \prod_{m=1}^{M} \prod_{j=1}^{C} \mathbf{P}(Y_m = j | \mathbf{X}, \mathbf{w})^{\mathbb{I}_{mj}} \\
&= \prod_{m=1}^{M} \prod_{j=1}^{C} \left(\frac{\exp\{\mathbf{w}_j^T \mathbf{x}_m\}}{\sum_{k=1}^{C} \exp\{\mathbf{w}_k^T \mathbf{x}_m\}} \right)^{\mathbb{I}_{mj}}
\end{aligned}
$$

Here $\mathbb{I}_{mj}$ equals 1 if the m^{th} data point belongs to class j, and equals 0 if the m^{th} data point does not belong to class j. Y_m is a random variable representing label of the m_{th} data point, and y_m is the label of the m_{th} data point. If we take the log and add the L2 regularization term, the above equation becomes $\boxed{L(\mathbf{w}_1, \cdots, \mathbf{w}_C) = \sum_{m=1}^{M} \sum_{j=1}^{C} \mathbb{I}_{mj} [\mathbf{w}_j^T \mathbf{x}_m - \ln \sum_{k=1}^{C} \exp\{\mathbf{w}_k^T \mathbf{x}_m\}] - \frac{\lambda}{2} ||\mathbf{w}_j||^2}$.

2. The expression for the j_{th} index is

$$
\begin{aligned}
\frac{\partial(\mathbf{w}_1, \cdots, \mathbf{w}_C)}{\partial \mathbf{w}_j)} &= \sum_{m=1}^{M} [\mathbb{I}_{mj} \mathbf{x}_m - \frac{\exp\{\mathbf{w}_j^T \mathbf{x}_m\}}{\sum_{k=1}^{C} \exp\{\mathbf{w}_k^T \mathbf{x}_m\}}] - \lambda \mathbf{w}_j \\
&= \boxed{\sum_{m=1}^{M} [\mathbb{I}_{mj} - \mathbf{P}(Y_m = j | \mathbf{X}, \mathbf{w})] \mathbf{x}_m - \lambda \mathbf{w}_j}.
\end{aligned}
$$

3. The update equation for weight vector $\mathbf{w}_j$ is $\boxed{\mathbf{w}_j + \eta \sum_{m=1}^{M} [\mathbb{I}_{mj} - \mathbf{P}(Y_m = j | \mathbf{X}, \mathbf{w})] \mathbf{x}_m - \eta \lambda \mathbf{w}_j}$.

4. The sequence of consecutive weight vectors will converge becasue the loss function itself is concave. It will converge when the loss function reaches its global maximum.

3 Feature Selection

1. The MLE estimate can be found by

$$\begin{aligned}\frac{\partial (Y-Xw)^T(Y-Xw)}{\partial w} &= -X^T(Y-X_w)=0\\ w &= (X^TX)^{-1}X^TY\\ w &= \boxed{[0.9484, -0.8811, 4.4696]}\end{aligned}$$

2. $\hat{w} =$

$$\begin{aligned}w &= (X^TX+\lambda I)^{-1}X^TY\\ w &= \boxed{[0.9029, -0.8715, 4.3416]}\end{aligned}$$

3. With *fminsearch* in MATLAB $w = \boxed{[0.9231, -0.8673, 4.4566]}$

4. After solving all 8 combinatorial cases, $w = \boxed{[0.9484, -0.8811, 4.4696]}$

5. The relation between the estimates of w in the four cases

   In the first case, the maximum likelihood estimates (MLE) aims to find the minimum value for the residual error without considering any assumptions or beliefs regarding w. However, with MLE being a consistent estimator, if the amount of data is small, the variance can be high. Typically MLE estimation is unbiased but has high variance. In the second case, the L_2 norm is an assumption that w is following a Gaussian distribution that has mean 0 and variance σ^2. With L_2 norm, if λ is a good value, it can help to avoid overfitting. In the ideal situation, irrelevant input should have weights set exactly to 0. In the third case, the L_1 norm is being penalized by decreasing w_1, w_2 and w_3 gradually down to zero. Those three parameters will be zeroed out if they become negative. L_1 norm can also be more computationally expensive than L_2 norm and Lasso is an efficient way of performing the L_1 regularization. In the fourth case, the L_0 norm is biased towards providing sparse solutions.

6. When $\lambda > 0$, we make a trade-off between minimizing the sum of squared errors and the magnitude of $\hat{w}$. In the following questions, we will explore this trade-off further. For the following, use the same data from data.mat.

   (a) The ratio of $||\hat{w}_{MLE}||_2^2/||Y - X\hat{w}_{MLE}||_2^2 = 21.6530/(1.9871e+03) = \boxed{0.0109}$

   (b) Doubling the number of training samples

      i. When N is doubled, $||Y - X\hat{w}_{MLE}||_2^2$ will also be doubled. When $N >> P$, this sum of squared erros depend directly on the number of training samples.

      ii. When you double the number of training samples, $||\hat{w}_{MLE}||_2^2$ should barely change. $||\hat{w}_{MLE}||_2^2$ does not depend directly on the number of training samples.

   (c) When $\boxed{\lambda = 3}$, $0.8 < ||\hat{w}||_2^2/||\hat{w}_{MLE}||_2^2 < 0.9$.

   (d) When $\boxed{\lambda = 19}$, $0.4 < ||\hat{w}||_2^2/||\hat{w}_{MLE}||_2^2 < 0.5$.

4 MDL on a toy dataset

1. Estimate the three linear regressions

   (a) The sum of square error

      i. $Err_1 = \boxed{460.0579}$.

      ii. $Err_2 = \boxed{300.6201}$.

      iii. $Err_3 = \boxed{300.5071}$.

   (b) 2 times the estimated bits to code the residual

      i. $ERR\_bits_1 = \boxed{182.1230}$.

      ii. $ERR\_bits_2 = \boxed{142.8351}$.

      iii. $ERR\_bits_3 = \boxed{142.8003}$.

   (c) 2 times the estimated bits to code each residual plus model under AIC

      i. $AIC\_bits_1 = \boxed{184.1230}$.

      ii. $AIC\_bits_2 = \boxed{146.8351}$.

      iii. $AIC\_bits_3 = \boxed{148.8003}$.

   (d) 2 times the estimated bits to code each residual plus model under BIC

      i. $BIC\_bits_1 = \boxed{188.1230}$.

      ii. $BIC\_bits_2 = \boxed{154.8351}$.

      iii. $BIC\_bits_3 = \boxed{160.8003}$.

2. Which model has the smallest minimum description length?

   (a) for AIC: $\boxed{\text{Model 2}}$.

   (b) for BIC: $\boxed{\text{Model 2}}$.

3. Test errors:

   (a) Model 1 test error = $\boxed{640.3078}$.

   (b) Model 2 test error = $\boxed{420.1459}$.

   (c) Model 3 test error = $\boxed{422.1606}$.