CIS 520: Problem Set #3

Due on October 1, 2017

Lyle Ungar

Eric Oh

Collaborators: Jiarui Lu

Problem 1

Naïve Bayes as a Linear Classifier

Solution

1. Note that using the given information, we have

$$P(x_i|y=1) = \alpha_i^{x_i}(1-\alpha_i)^{1-x_i}$$
$$P(x_i|y=-1) = \beta_i^{x_i}(1-\beta_i)^{1-x_i}$$

Then it follows that

$$\begin{aligned} P(\mathbf{x}|y=1) &= \prod_{i=1}^{n} P(x_i|y=1) \qquad \text{(by the Naive Bayes assumption)} \\ &= \prod_{i=1}^{n} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i} \\ P(\mathbf{x}|y=-1) &= \prod_{i=1}^{n} P(x_i|y=-1) \qquad \text{(by the Naive Bayes assumption)} \\ &= \prod_{i=1}^{n} \beta_i^{x_i}(1-\beta_i)^{1-x_i} \end{aligned}$$

2. Let A_j denote a random variable such that

$$A_j = \left\{ \begin{array}{ll} 1 & \text{if } y_j = 1 \\ 0 & \text{if } y_j = -1 \end{array} \right\}$$

The likelihood is given by:

$$\begin{aligned} L(p, \alpha_i, \beta_i) &= \prod_{j=1}^{m} P(x_{ji}, y_j) = \prod_{j=1}^{m} P(y_j)P(x_{ji}|y_j) = \prod_{j=1}^{m} P(y_j) \prod_{i=1}^{n} P(x_{ji}|y_j) \\ &= \prod_{j=1}^{m} \left[p \prod_{i=1}^{n} \alpha_i^{x_{ji}}(1-\alpha_i)^{x_{ji}} \right]^{A_j} \left[(1-p) \prod_{i=1}^{n} \beta_i^{x_{ji}}(1-\beta_i)^{x_{ji}} \right]^{1-A_j} \\ l(p, \alpha_i, \beta_i) = \log L(p, \alpha_i, \beta_i) &= \sum_{j=1}^{m} \left[A_j \left\{ \log(p) + \sum_{i=1}^{n} x_{ji} \log(\alpha_i) + (1-x_{ji}) \log(1-\alpha_i) \right\} \right. \\ &\quad \left. + (1-A_j) \left\{ \log(1-p) + \sum_{i=1}^{n} x_{ji} \log(\beta_i) + (1-x_{ji}) \log(1-\beta_i) \right\} \right] \end{aligned}$$

Problem 1 continued on next page...

The MLEs are found as follows:

$$\frac{dl}{dp} = \sum_{j=1}^{m}\left[\frac{A_j}{p} - \frac{(1-A_j)}{1-p}\right] \equiv 0 \quad \Rightarrow \quad \hat{p} = \frac{\sum_{j=1}^{m} A_j}{m}$$

$$\frac{dl}{d\alpha_i} = \sum_{j=1}^{m}\left[A_j \sum_{i=1}^{n}\frac{x_{ji}}{\alpha_i} - A_j \sum_{i=1}^{n}\frac{1-x_{ji}}{1-\alpha_i}\right] \equiv 0$$

$$\Downarrow$$

$$\sum_{i=1}^{n}\frac{1}{\alpha_i}\sum_{j=1}^{m} A_j x_{ji} = \sum_{i=1}^{n}\frac{1}{1-\alpha_i}\sum_{j=1}^{m} A_j(1-x_{ji}) \quad \Rightarrow \hat{\alpha_i} = \frac{\sum_j A_j x_{ji}}{\sum_j A_j}$$

$$\frac{dl}{d\beta_i} = \sum_{j=1}^{m}\left[(1-A_j)\sum_{i=1}^{n}\frac{x_{ji}}{\beta_i} - (1-A_j)\sum_{i=1}^{n}\frac{1-x_{ji}}{1-\beta_i}\right] \equiv 0$$

$$\Downarrow$$

$$\sum_{i=1}^{n}\frac{1}{\beta_i}\sum_{j=1}^{m}(1-A_j)x_{ji} = \sum_{i=1}^{n}\frac{1}{1-\beta_i}\sum_{j=1}^{m}(1-A_j)(1-x_{ji}) \quad \Rightarrow \hat{\beta_i} = \frac{\sum_j (1-A_j)x_{ji}}{\sum_j (1-A_j)}$$

The second derivatives of the three derivatives above are all easily shown to be negative (not shown for the sake of brevity).

3. We are given that the NB classifier is $h(x) = \text{argmax}_{y\in\pm 1}\hat{P}(y|x)$ where

$$\hat{P}(y|x) \propto \hat{P}(x|y)\hat{P}(y)$$

From this definition, we would classify y_i as 1 if

$$\hat{P}(y_i = 1|\mathbf{x}) > \hat{P}(y_i = -1|\mathbf{x})$$
$$\hat{P}(y_i = 1|\mathbf{x}) - \hat{P}(y_i = -1|\mathbf{x}) > 0$$

and as -1 otherwise. Thus our NB classifier can be written as

$$h(x) = \text{sign}(\hat{P}(y_i = 1|\mathbf{x}) - \hat{P}(y_i = -1|\mathbf{x}))$$

4. Consider the quantity $\hat{P}(y_i = 1|\mathbf{x}) - \hat{P}(y_i = -1|\mathbf{x})$.

$$\begin{aligned}
\hat{P}(y_i = 1|\mathbf{x}) - \hat{P}(y_i = -1|\mathbf{x}) &= \hat{p}\prod_{i=1}^{n}\hat{\alpha}_i^{x_i}(1-\hat{\alpha}_i)^{x_i} - (1-\hat{p})\prod_{i=1}^{n}\hat{\beta}_i^{x_i}(1-\hat{\beta}_i)^{x_i} && \text{(by Bayes' Rule)}\\
&= \log(\hat{p}) + \sum_{i=1}^{n} x_i\log(\hat{\alpha}_i) + (1-x_i)\log(1-\hat{\alpha}_i)\\
&\quad - \log(1-\hat{p}) + \sum_{i=1}^{n} x_i\log(\hat{\beta}_i) + (1-x_i)\log(1-\hat{\beta}_i) && \text{(taking logs)}\\
&= \log\frac{\hat{p}}{1-\hat{p}} + \sum_{i=1}^{n}\log\left(\frac{1-\hat{\alpha}_i}{1-\hat{\beta}_i}\right) + \sum_{i=1}^{n} x_i\log\left(\frac{\hat{\alpha}_i(1-\hat{\beta}_i)}{(1-\hat{\alpha}_i)\hat{\beta}_i}\right)
\end{aligned}$$

Note that since the logarithm is an increasing function and all of our MLEs are positive values, taking

logs does not change the argument to the sign function for the NB classifier. Now we can let

$$b = \log \frac{\hat{p}}{1-\hat{p}} + \sum_{i=1}^{n} \log\left(\frac{1-\hat{\alpha}_i}{1-\hat{\beta}_i}\right)$$
$$w_i = \sum_{i=1}^{n} \log\left(\frac{\hat{\alpha}_i(1-\hat{\beta}_i)}{(1-\hat{\alpha}_i)\hat{\beta}_i}\right)$$

and it follows that the NB classifier can be written $h(x) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$.

Problem 2

Multiclass Logistic Regression

Solution

1. Let y_i^j denote a random variable such that

$$y_i^j = \left\{ \begin{array}{ll} 1 & \text{if } y_i = j \\ 0 & \text{if } y_i \neq j \end{array} \right\}$$

for $j \in \{1, 2, \ldots, C\}$. The likelihood is then given by

$$\begin{aligned} L(\mathbf{w}_1, \ldots, \mathbf{w}_c) &= \prod_{i=1}^{M} L_i(\mathbf{w}_1, \ldots, \mathbf{w}_c) && \text{(product over all data points)} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{C} P(y_i = j|\mathbf{x}; \mathbf{w}_1, \ldots, \mathbf{w}_c)^{y_i^j} && \text{(by given probability and independence assumption)} \\ l(\mathbf{w}_1, \ldots, \mathbf{w}_c) &= \sum_{i=1}^{M}\sum_{j=1}^{C} y_i^j \log P(y_i = j|\mathbf{x}; \mathbf{w}_1, \ldots, \mathbf{w}_c) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{C} y_i^j \left[\mathbf{w}_j^T\mathbf{x} - \log\sum_{k=1}^{C}\exp(\mathbf{w}_k^T\mathbf{x})\right] && \text{(taking the log)} \end{aligned}$$

To add an L_2 regularization term, we add a Gaussian prior where $w \sim \text{N}(0, \lambda^{-1})$. Note that this simplifies to

$$f(w;\lambda) = \frac{1}{\sqrt{2\pi\lambda^{-1}}}\exp\left(-\frac{\lambda}{2}w^2\right) \propto \exp\left(-\frac{\lambda}{2}w^2\right) \qquad \text{(likelihood is a function of w)}$$

Thus the L_2 penalized likelihood is given by

$$\begin{aligned} L(\mathbf{w}_1, \ldots, \mathbf{w}_c) &= \prod_{i=1}^{M}\prod_{j=1}^{C} \left\{P(y_i = j|\mathbf{x}; \overrightarrow{w}) f(\overrightarrow{w}_j; \lambda^{-1})\right\}^{y_i^j} \\ &\propto \prod_{i=1}^{M}\prod_{j=1}^{C} \left\{P(y_i = j|\mathbf{x}; \overrightarrow{w})\exp\left(-\frac{\lambda}{2}w_j^2\right)\right\}^{y_i^j} \\ l(\mathbf{w}_1, \ldots, \mathbf{w}_c) &= \sum_{i=1}^{M}\sum_{j=1}^{C} y_i^j \left[\mathbf{w}_j^T\mathbf{x} - \log\sum_{k=1}^{C}\exp(\mathbf{w}_k^T\mathbf{x}) - \frac{\lambda}{2}w_j^2\right] \end{aligned}$$

Problem 2 continued on next page...

2.

$$\begin{aligned}\frac{dl}{dw_j} &= \sum_{i=1}^{M} y_i^j \mathbf{x} - y_i^j \frac{\mathbf{x}\exp(\mathbf{w}_j^T \mathbf{x})}{\sum_{k=1}^{C} \exp(\mathbf{w}_j^T \mathbf{x})} - y_i^j \lambda w_j \\ &= \sum_{i=1}^{M} y_i^j \mathbf{x}\,(1 - P(y_i = j|\mathbf{x}; \mathbf{w})) - y_i^j \lambda w_j\end{aligned}$$

3.

$$w_j^{t+1} = w_j^t + \eta \left[\sum_{i=1}^{M} y_i^j \mathbf{x}\,(1 - P(y_i = j|\mathbf{x}; \mathbf{w})) - y_i^j \lambda w_j \right]$$

4. The weights will converge to a local maximum because the log-likelihood is a convex upwards function. This local maximum will be the point where the gradients are 0.

Problem 3

Feature Selection

Solution

1. Assuming a i.i.d data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the MLE is found as follows:

$$\begin{aligned}L(w) &= \prod_{i=1}^{n} (\sqrt{2\pi})^{-\frac{1}{2}} \sigma^{-1} \exp\left[-\frac{1}{2\sigma^2} (y_i - w x_i)^2 \right] \\ l(w) &= \log L(w) \propto \sum_{i=1}^{n} -\frac{1}{2\sigma^2} (y_i - w x_i)^2 \\ \frac{dl}{dw} &= \sum_{i=1}^{n} \frac{1}{\sigma^2} x_i (y_i - w x_i) \equiv 0 \qquad \Rightarrow \qquad \hat{w}_{\text{MLE}} = (X^T X)^{-1} X^T Y\end{aligned}$$

Using the provided dataset (data.mat), the numerical MLE is given by

$$\hat{w}_{\text{MLE}} = \begin{bmatrix} 0.8891 \\ -0.826 \\ 4.1902 \end{bmatrix}$$

2.

$$\hat{w} = \begin{bmatrix} 0.8646 \\ -0.8210 \\ 4.1218 \end{bmatrix}$$

3.

$$\hat{w} = \begin{bmatrix} 0.8749 \\ -0.8182 \\ 4.1829 \end{bmatrix}$$

4.

$$\hat{w} = \begin{bmatrix} 0.8891 \\ -0.826 \\ 4.1902 \end{bmatrix}$$

Problem 3 continued on next page...

5. The MLE follows from the standard linear regression theory. Adding a regularization term shrinks the MLE coefficients corresponding to different penalties. The L_2 regularization shrinks the coefficients according to the L_2 norm, meaning all coefficients are smaller but none are set to 0. We see that due to the penalty being so small and the squared error dominating, there is not much shrinkage. L_1 regularization shrinks coefficients according to the L_1 norm, meaning that some coefficients are shrunk to 0. However, our penalty is small and the squared error dominates again, meaning there is some shrinkage but none of the coefficients are 0. L_0 regularization shrinks coefficients according to the L_0 norm, meaning the penalty is applied to the number of non-zero features. Again, our penalty is small and the number of features is small so in this case we get exactly the MLE.

6. (a) 0.0061
   (b) (i) When going from N to 2N, the sum of squared error will clearly increase as N increases as we sum over more subjects. However, if we divide the sum of squared error by the number of subjects, consistency results give us that the quantity would converge to the true variance of the outcome.
   (ii) We would not expect $\hat{w}_{\text{MLE}}$ to change much as N increases to 2N. From consistency results, it would converge to the true parameter of the data generating distribution as N increases.
   (c) $\lambda = 5$ yields a ratio of 0.8523.
   (d) $\lambda = 26$ yields a ratio of 0.4896

Problem 4

Entropy and Minimum Description Length

Solution

1. The minimum number of bits needed to code the sequence is given by the entropy:

$$\text{Entropy} = -\left[\frac{3}{16}\log\left(\frac{3}{16}\right) + \frac{13}{16}\log\left(\frac{13}{16}\right)\right] = 0.6962$$

2. The original penalty for the RIC is $\log_2 f = -\log_2\left(\frac{1}{f}\right)$ based on the prior belief of one feature being included in the model. If we had prior belief that $\frac{3}{16}$ features would be selected for the model, we could update the penalty to be

$$\lambda = -\log_2\left(\frac{3}{16}\right) = \log_2\left(\frac{16}{3}\right)$$

Problem 5

MDL on a toy dataset

Solution

1. (a) (i) $\text{Err}_1 = 1277.9$
   (ii) $\text{Err}_2 = 835.06$
   (iii) $\text{Err}_3 = 834.74$
   (b) (i) $\text{Err}_{\text{bits}_1} = 552.91$
   (ii) $\text{Err}_{\text{bits}_2} = 474.33$

Problem 5 continued on next page...

(iii) $\text{Err}_{\text{bits}_1} = 474.26$

(c) (i) $\text{AIC}_{\text{bits}_1} = 554.91$

(ii) $\text{AIC}_{\text{bits}_2} = 478.33$

(iii) $\text{AIC}_{\text{bits}_1} = 480.26$

(d) (i) $\text{BIC}_{\text{bits}_1} = 558.91$

(ii) $\text{BIC}_{\text{bits}_2} = 486.33$

(iii) $\text{BIC}_{\text{bits}_1} = 492.26$

2. (a) $y_2 = w_1 x_1 + w_2 x_2$

(b) $y_2 = w_1 x_1 + w_2 x_2$

3. Yes, the errors on the test set are higher overall but the second model with two features is still the best model in terms of minimum description length for AIC and BIC.