CIS 520, Machine Learning, Fall 2018: Assignment 3

Simran Arora, Shubhankar Patankar

1 Naïve Bayes as a Linear Classifier

1. Conditional probability of $\mathbf{x}$ given y,

$$\Pr(\mathbf{x}|y=1) = \prod_{i=1}^{n} \Pr(x_i|y=1)$$

$\Pr(x_i|y=1) = \alpha_i$ if $x_i = 1$ or $(1-\alpha_i)$ if $x_i = 0$

$$\begin{aligned}\therefore \Pr(x_i|y=1) &= (\alpha_i)^{I[x_i=1]}(1-\alpha_i)^{I[x_i=0]}\\ &= (\alpha_i)^{I[x_i=1]}(1-\alpha_i)^{1-I[x_i=1]}\\ &= (\alpha_i)^{x_i}(1-\alpha_i)^{1-x_i}\end{aligned}$$

$$\therefore \Pr(\mathbf{x}|y=1) = \prod_{i=1}^{n} (\alpha_i)^{x_i}(1-\alpha_i)^{1-x_i}$$

Similarly,

$$\therefore \Pr(\mathbf{x}|y=-1) = \prod_{i=1}^{n} (\beta_i)^{x_i}(1-\beta_i)^{1-x_i}$$

2. For the MLE of the parameters:
To solve for $\widehat{p}$ our goal is to maximize the log-likelihood of observing the data ($argmax_p \sum_{i=1}^{m} logp(x_i|y)$).
We know that $P(y=1) = p$. This gives us the problem:

$$argmax_p \sum_{i=1}^{m} \left(\frac{1+y_i}{2}\right) log[(p)Pr(x_i|y_i=1)] + \left(\frac{1-y_i}{2}\right) log[(1-p)Pr(x_i|y_i=-1)]$$

Taking the derivative with respect to p:

$$\frac{d}{dp}\left\{\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right) log[(p)Pr(x_i|y_i=1)] + \left(\frac{1-y_i}{2}\right) log[(1-p)Pr(x_i|y_i=-1)]\right\}\right\}$$

$$= \sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)\frac{1}{(p)Pr(x_i|y_i=1)}Pr(x_i|y_i=1) + \left(\frac{1-y_i}{2}\right)\frac{1}{(1-p)Pr(x_i|y_i=-1)}(-Pr(x_i|y_i=-1))\right\}$$

$$= \sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)\frac{1}{p} - \left(\frac{1-y_i}{2}\right)\frac{1}{(1-p)}\right\}$$

Set this equal to 0:

$$\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)\frac{1}{p}-\left(\frac{1-y_i}{2}\right)\frac{1}{(1-p)}\right\}=0$$

$$(1-p)\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}\left(\frac{1-y_i}{2}\right)=0$$

$$\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}\left(\frac{1-y_i}{2}\right)=0$$

$$\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)+\left(\frac{1-y_i}{2}\right)\right\}=0$$

$$\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}\left(\frac{1}{2}+\frac{1}{2}\right)=0$$

$$\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)\sum_{i=1}^{m}1=0$$

$$\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)-(p)(m)=0$$

$$\frac{\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)}{m}=p$$

We know that$\sum_{i=1}^{m}(\frac{1+y_i}{2})$ equals the number of occasions on which $y_i = 1$ because if $y_i = 1$, $\frac{1+y_i}{2} = 1$ and if $y_i = 0$, $\frac{1+y_i}{2} = 0$. Thus, the $\widehat{p} = \frac{\sum_{i=1}^{m}(\frac{1+y_i}{2})}{m}$ means that $\widehat{p}$ is given by the number of occurrences of $y_i = 1$ divided by the total number of data points, which intuitively makes sense because $p = Pr(y_i = 1)$.

To solve for $\widehat{\alpha_i}$, our goal is to maximize the log-likelihood of observing the data with respect to α_i.

$$\frac{d}{d\alpha_i}\left\{\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)log[(p)Pr(x_i|y_i=1)]+\left(\frac{1-y_i}{2}\right)log[(1-p)Pr(x_i|y_i=-1)]\right\}\right\}$$

$$=\frac{d}{d\alpha_i}\left\{\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)log\left[p\left(\prod_{j=1}^{n}\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}\right)\right]+\left(\frac{1-y_i}{2}\right)log\left[(1-p)\left(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j})\right)\right]\right\}\right\}$$

$$=\frac{d}{d\alpha_i}\left\{\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)\left[log(p)+\sum_{j=1}^{n}log(\alpha_j^{x_j}(1-\alpha_j)^{1-x_j})\right]+\left(\frac{1-y_i}{2}\right)log\left[(1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j})\right]\right\}\right\}$$

$$=\frac{d}{d\alpha_i}\left\{\sum_{i=1}^{m}\left\{\left(\frac{1+y_i}{2}\right)\left[log(p)+\sum_{j=1}^{n}(x_jlog(\alpha_j)+(1-x_j)log(1-\alpha_j))\right]+\left(\frac{1-y_i}{2}\right)log\left[(1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j})\right]\right\}\right\}$$

$$=\sum_{i=1}^{m}\left(\frac{1+y_i}{2}\right)\left[\sum_{j=1}^{n}\frac{d}{d\alpha_i}\left(x_jlog(\alpha_j)+(1-x_j)log(1-\alpha_j)\right)\right]$$

$$= \sum_{i=1}^{m} \left(\frac{1+y_i}{2}\right) \left\{ \sum_{j=1}^{n} \left[\left(\frac{x_j}{\alpha_i}\right) - \left(\frac{1-x_j}{1-\alpha_i}\right) \right] \right\}$$

Setting this derivative equal to 0, we get:

$$\sum_{i=1}^{m} \left(\frac{1+y_i}{2}\right) \left[\sum_{j=1}^{n} \left(\frac{x_j}{\alpha_i}\right) - \sum_{j=1}^{n} \left(\frac{1-x_j}{1-\alpha_i}\right) \right] = 0$$

$$(1-\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \Bigg) = (\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} \Big(1 - \sum_{j=1}^{n} x_j\Big)\Bigg)$$

$$\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \Bigg) - (\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \Bigg) = (\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} 1 \Bigg) - (\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \Bigg)$$

$$\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \Bigg) = (\alpha_i)\Bigg(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} 1 \Bigg)$$

$$\therefore \widehat{\alpha_i} = \frac{\big(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} x_j \big)}{\big(\sum_{i=1}^{m} \frac{1+y_i}{2} \sum_{j=1}^{n} 1 \big)}$$

This intuitively means that $\widehat{\alpha_i}$ is the count of observations in class $y = 1$ with attribute $x_i = 1$ divided by the total count of observations in class $y = 1$.

By symmetry to the above,

$$\therefore \widehat{\beta_i} = \frac{\big(\sum_{i=1}^{m} \frac{1-y_i}{2} \sum_{j=1}^{n} x_j \big)}{\big(\sum_{i=1}^{m} \frac{1-y_i}{2} \sum_{j=1}^{n} 1 \big)}$$

$\widehat{\beta_i}$ is the count of observations in class $y = -1$ with attribute $x_i = 1$ divided by the total count of observations in class $y = -1$.

3. $h(x)$ can be written as:

$$h(\vec{x}) = argmax_{y \in \{\pm 1\}} \widehat{Pr}(y|\vec{x})$$

This returns the y class value that yields the highest probability for $\widehat{Pr}(y|\vec{x})$. Want to prove that this is equivalent to the function $h'(\vec{x}) = sign(\widehat{Pr}(1|\vec{x}) - \widehat{Pr}(-1|\vec{x})$.

If $h(\vec{x})$ returns $y = 1$, then this means that $\widehat{Pr}(y = 1|\vec{x}) > \widehat{Pr}(y = -1|\vec{x})$. Let $x = \widehat{Pr}(y = 1|\vec{x}) - \widehat{Pr}(y = -1|\vec{x})$ in which $x > 0$. Then, $h'(\vec{x}) = sign(x)$ which thus returns +1. Meanwhile, if $h(\vec{x})$ returns $y = -1$, then this means that $\widehat{Pr}(y = -1|\vec{x}) > \widehat{Pr}(y = +1|\vec{x})$. Let $x = \widehat{Pr}(y = -1|\vec{x}) - \widehat{Pr}(y = +1|\vec{x})$ in which $x < 0$. Then, $h'(\vec{x}) = sign(x)$ which thus returns -1. Thus we can see that $h(\vec{x})$ and $h'(\vec{x})$ always return the same result, and thus they are equivalent.

4. **w** and b from $h(x)$: We want to show that $h(\vec{x}) = sign(\vec{w}^T \vec{x} + b)$. First by first using Bayes rule and then the total probability formula:

$$\widehat{Pr}(y = a|\vec{x}) = \frac{Pr(\vec{x}|y = a)Pr(y = a)}{Pr(\vec{x})}$$

We want the a class value that maximizes the probability:

$$argmax_y \widehat{Pr}(y=a|\vec{x}) = argmax_y \frac{Pr(\vec{x}|y=a)Pr(y=a)}{Pr(\vec{x})}$$

Since the denominator of this does not depend on y, this is equivalent to:

$$argmax_y Pr(\vec{x}|y=a)Pr(y=a)$$

From equation 2 we can rearrange to get:

$$log(Pr(\vec{x}|y=1)) + log(Pr(y=1)) - log(Pr(\vec{x}|y=-1)) - log(Pr(y=-1))$$

Because the activation of $log(\alpha_i)$ vs $log(\beta_i)$ depends on the value of $\vec{x}$, whether it's positive or 0, we can rearrange this to be:

$$(log(\alpha_i) - log(\beta_i))^T \vec{x} + (log(p) - log(1-p))$$

Thus, we get that $\vec{w} = (log(\alpha_i) - log(\beta_i))$ and $b = (log(p) - log(1-p))$.

2 Multiclass Logistic Regression

1. For multi-class classification, assuming there are C different classes, for each class,

$$\mathbf{P}(C_j \mid X = \mathbf{x}) = \frac{\exp\{\mathbf{w}_j^T\mathbf{x}\}}{\sum_{k=1}^{C}\exp\{\mathbf{w}_k^T\mathbf{x}\}} \quad \forall j \in \{1, 2, .., C\}$$

Let the training data have N points and P features for each point. The target matrix T has dimensions $N \times P$. Using a 1-of-C coding scheme, each row of T has zeros in all positions except the column corresponding to the right class. The probability of observing the matrix T is the probability of observing each of its elements. Therefore, it can be written as follows:

$$\begin{aligned}\mathbf{P}(T \mid \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C, \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N) = {} & \mathbf{P}(C_1 \mid \mathbf{x}_1) \ldots \mathbf{P}(C_C \mid \mathbf{x}_1) \ldots \\ & \mathbf{P}(C_1 \mid \mathbf{x}_2) \ldots \mathbf{P}(C_C \mid \mathbf{x}_2) \ldots \\ & \ldots \\ & \mathbf{P}(C_1 \mid \mathbf{x}_N) \ldots \mathbf{P}(C_C \mid \mathbf{x}_N)\end{aligned}$$

Each term in the product above has the appropriate weight vector as part of the argument that has been dropped for clarity.

$$\therefore \mathbf{P}(T \mid \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C, \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N) = \prod_{n=1}^{N}\prod_{j=1}^{C} \mathbf{P}(C_j \mid \mathbf{x}_n)^{t_{nj}}$$

The exponent t_{nj} is added to reflect the 1-of-C coding scheme. Each row has only one position that is non-zero. For instance, assuming that the first data point belongs to class 3, $t_{13} = 1$ and $t_{1,k\neq 3} = 0$. The terms in the product that have 0 as the exponent equal 1 and do not contribute to the likelihood. The log likelihood can then be written as follows:

$$L(\mathbf{w}_1, ..., \mathbf{w}_C) = \sum_{n=1}^{N}\sum_{j=1}^{C}(t_{nj})log\big[\mathbf{P}(C_j \mid \mathbf{x}_n)\big]$$

With the logarithm the products change into summations and the exponent drops to the front. Adding an L2 regularization term, in which each weights vector is penalized the same amount λ, gives:

$$\begin{aligned}
L(\mathbf{w}_1, ..., \mathbf{w}_C) &= \sum_{n=1}^{N}\sum_{j=1}^{C}(t_{nj})log\big[\mathbf{P}(C_j \mid \mathbf{x}_n)\big] - \lambda\sum_{j=1}^{C}\|\mathbf{w}_j\|_2^2 \\
&= \sum_{n=1}^{N}\sum_{j=1}^{C}(t_{nj})log\left[\frac{\exp\{\mathbf{w}_j^T\mathbf{x}_n\}}{\sum_{k=1}^{C}\exp\{\mathbf{w}_k^T\mathbf{x}_n\}}\right] - \lambda\sum_{j=1}^{C}\|\mathbf{w}_j\|_2^2 \\
&= \sum_{n=1}^{N}\sum_{j=1}^{C}(t_{nj})\mathbf{w}_j^T\mathbf{x}_n - \sum_{n=1}^{N}\sum_{j=1}^{C}(t_{nj})log\left[\sum_{k=1}^{C}\exp(\mathbf{w}_k^T\mathbf{x}_n)\right] - \lambda\sum_{j=1}^{C}\|\mathbf{w}_j\|_2^2
\end{aligned}$$

2. Differentiating with respect to $\mathbf{w}_j$,

$$\frac{\partial L(\mathbf{w}_1, ..., \mathbf{w}_C)}{\partial \mathbf{w}_j} = \sum_{n=1}^{N} t_{nj}\mathbf{x}_n - \sum_{n=1}^{N}\left[\frac{\exp(\mathbf{w}_j^T\mathbf{x}_n)}{\sum_{k=1}^{C}\exp\{\mathbf{w}_k^T\mathbf{x}_n\}}\right]\mathbf{x}_n - 2\lambda\mathbf{w}_j$$

   The second term in the derivative follows from the observation that in terms where $k \neq j$, the chain rule causes the term to disappear. This only leaves the $k = j$ case to differentiate according to the chain rule.

$$\frac{\partial L(\mathbf{w}_1, ..., \mathbf{w}_C)}{\partial \mathbf{w}_j} = \sum_{n=1}^{N}\left[t_{nj} - \left(\frac{\exp(w_j^T\mathbf{x}_n)}{\sum_{k=1}^{C}\exp\{\mathbf{w}_k^T\mathbf{x}_n\}}\right)\right]\mathbf{x}_n - 2\lambda\mathbf{w}_j$$

3. The update equation for the weights vector $\mathbf{w}_j$ can be written as follows:

$$\begin{aligned}
\mathbf{w}_{j+1} &= \mathbf{w}_j + \eta\frac{\partial L(\mathbf{w}_1, ..., \mathbf{w}_C)}{\partial \mathbf{w}_j} \\
&= \mathbf{w}_j + \eta\left\{\sum_{n=1}^{N}\left[t_{nj} - \left(\frac{\exp(w_j^T\mathbf{x}_n)}{\sum_{k=1}^{C}\exp\{\mathbf{w}_k^T\mathbf{x}_n\}}\right)\right]\mathbf{x}_n - 2\lambda\mathbf{w}_j\right\}
\end{aligned}$$

4. The Hessian for the likelihood function is positive definite implying that the sequence of consecutive weight vectors converge.

3 Feature Selection

$$\widehat{w} = \arg\min_{w} \quad \|Y - Xw\|_2^2 + \lambda\|w\|_0$$

1. The MLE estimate for $\widehat{w}_{MLE}$,

$$\begin{aligned}\frac{\partial \widehat{w}}{\partial w} &= \frac{\partial\big[(Y-Xw)^T(Y-Xw)\big]}{\partial w}\\ &= \frac{\partial\big[(Y^T-w^TX^T)(Y-Xw)\big]}{\partial w}\\ &= \frac{\partial\big[(Y^TY-w^TX^TY-w^TX^TY+w^TX^TXw)\big]}{\partial w}\\ &= \frac{\partial\big[(Y^TY-2w^TX^TY+w^TX^TXw)\big]}{\partial w}\\ &= -2X^TY+2X^TXw\end{aligned}$$

Setting equal to zero,

$$-2X^TY+2X^TXw=0$$

$$\therefore X^TX\widehat{w}_{MLE}=X^TY$$

$$\widehat{w}_{MLE}=(X^TX)^{-1}X^TY$$

For the given dataset, $\widehat{w}_{MLE}=[0.9484;-0.8811;4.4696]$

2. $\lambda=1$ for L2-penalty,

$$\widehat{w}_{MLE}=[0.9029;-0.8715;4.3416]$$

3. $\lambda=1$ for L1-penalty,

$$\widehat{w}_{MLE}=[0.9231;-0.8673;4.4565]$$

4. $\lambda=1$ for L0-penalty, The eight combinatorial cases for the elements of the weight vector to be zero are the following: $(0;0;0)_1,(1;0;0)_2,(0;1;0)_3,(0;0;1)_4,(1;1;0)_5,(1;0;1)_6,(0;1;1)_7,(1;1;1)_8$. 1 is used as a stand-in for some non-zero weight value. For each element of the weight vectors that is zero, the corresponding feature in the training data has no influence on the label. The eight optimized weight vectors computed are as follows:

$$\widehat{w}_{MLE_1}=(0;0;0)$$

$$\widehat{w}_{MLE_2}=(1.2370;0;0)$$

$$\widehat{w}_{MLE_3}=(0;-1.6033;0)$$

$$\widehat{w}_{MLE_4}=(0;0;4.5794)$$

$$\widehat{w}_{MLE_5}=(0.9629;-1.4856;0)$$

$$\widehat{w}_{MLE_6}=(1.1084;0;4.5628)$$

$$\widehat{w}_{MLE_7}=(0;-0.9967;4.4712)$$

$$\widehat{w}_{MLE_8}=(0.9485;-0.8810;4.4696)$$

The cost function is the lowest when all three features are included in the model (case 8).

$$\therefore \widehat{w}_{MLE}=(0.9485;-0.8810;4.4696)$$

5. Ridge regression shrinks the weights closer to 0 compared to L0. The applied shrinkage is proportional to how far the weights are from 0. As a result, bigger weights are shrunk more. In general, it can be seen that the L2 penalty causes the MLE weights to be shrunk to values closer to 0 compared to the L0 and L1 penalties. L0 shrinks all weights within a threshold to 0 leaving others unchanged. In our case, the penalty λ is not large enough to drive any of the weights to 0. Lasso shrinks all weights within a threshold to 0 and shrinks the others by a constant amount. For the given data, the penalty is not large enough to drive any of the weights to 0 for lasso. Based on this, it is reasonable that the weights under L2 are closest to 0 and the L1 weights are generally between L0 and L2. The MLE estimate of the weights is the largest since there is no regularization penalty that causes shrinkage.

6. Trade-off between minimizing the SSE and the magnitude of $\widehat{w}$.

   (a)
   $$||\widehat{w}_{MLE}||_2^2 \;/\; ||Y - X\widehat{w}_{MLE}||_2^2 = 0.0109$$
   .

   (b) Effect of adding data points:

      i. When more samples are added to the dataset, the SSE increases. Each new error term is squared and added to the cumulative sum of the other terms, causing the SSE to grow with the number of samples. Additionally, in any realistic data set with inherent noise, even the best prediction causes the square of the noise to be added to the SSE.

      ii. The sum of the squared weights does not significantly change when more samples are introduced unless the number of samples to begin with is low. For instance, if 2 samples are added to a data set that previously contained only 1 sample, the weights for this new fit would substantially differ from the case with just the one data point. On the other hand, if the dataset has a large number of data points to begin with, adding twice as many samples might cause negligibly small changes in the sum of the squared weights. This presumes that the new samples do not have a large number of outliers.

   (c) λ such that $0.8 < ||\widehat{w}||_2^2 \;/\; ||\widehat{w}_{MLE}||_2^2 < 0.9$.
   $$\lambda = 3$$

   (d) λ such that $0.4 < ||\widehat{w}||_2^2 \;/\; ||\widehat{w}_{MLE}||_2^2 < 0.5$.
   $$\lambda = 16$$

4 MDL

1. Estimate the three linear regressions

$$\begin{aligned} y_1 &= w_1x_1 \\ y_2 &= w_1x_1 + w_2x_2 \\ y_3 &= w_1x_1 + w_2x_2 + w_3x_3 \end{aligned}$$

   (a) the sum of square error
   i) $\text{Err}_1 = 460.0579$
   ii) $\text{Err}_2 = 300.6201$
   iii) $\text{Err}_3 = 300.5071$

(b) 2 times the estimated bits to code the residual ($n \log \frac{Error}{n}$)
i) $\text{ERR\_bits}_1 = 182.1230$
ii) $\text{ERR\_bits}_2 = 142.8351$
iii) $\text{ERR\_bits}_3 = 142.8003$

(c) 2 times the estimated bits to code each residual plus model under AIC ($2 * 1$ bit to code each feature)
i) $\text{AIC\_bits}_1 = 184.1230$
ii) $\text{AIC\_bits}_2 = 146.8351$
iii) $\text{AIC\_bits}_3 = 148.8003$

(d) 2 times the estimated bits to code each residual plus model under BIC ($2 * (1/2)log(n)$ bits to code each feature)
i) $\text{BIC\_bits}_1 = 188.1230$
ii) $\text{BIC\_bits}_2 = 154.8351$
iii) $\text{BIC\_bits}_3 = 160.8003$

2. Which model has the smallest minimum description length?
a) for AIC: Model 2
b) for BIC: Model 2

3. Included in the kit is a test data set; does the error on the test set for the three models correspond to what is expected from MDLs? Please compute and show the test errors and briefly explain it in one sentence.
i) Test $\text{Error}_1 = 640.3078$
ii) Test $\text{Error}_2 = 420.1459$
iii) Test $\text{Error}_3 = 422.1606$
The test error corresponds to what is expected from the MDLs since the MDL for Model 2 is the smallest.