

Name: Elise McElhiney Collaborators: Bekah Coggin, David Young, David Menager

1. 4.18 Naive Bayes with mixed features

Consider a 3-class naive Bayes classifier with one binary feature and one Gaussian feature:

$$y \sim \text{Mu}(y|\pi, 1), x_1|y = c \sim \text{Ber}(x_1|\theta_c), x_2|y = c \sim \mathcal{N}(x_2|\mu_c, \sigma_c^2)$$

Let the parameter vectors be as follows:

$$\boldsymbol{\pi} = (0.5, 0.25, 0.25), \quad \boldsymbol{\theta} = (0.5, 0.5, 0.5), \quad \boldsymbol{\mu} = (-1, 0, 1), \quad \boldsymbol{\sigma}^2 = (1, 1, 1)$$

Solution:

$\circ \triangleq$ Hadamard Product

a. Compute $p(y|x_1 = 0, x_2 = 0)$ (the result should be a vector of 3 numbers that sums to 1)

$$p(y|x_1 = 0, x_2 = 0) = \frac{\boldsymbol{\pi} \circ \boldsymbol{\theta} \circ N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}{\sum_j \pi_j \theta_j N(x_j|\mu_j, \sigma_j^2)} = \frac{(0.06, 0.05, 0.03)}{0.14} = (0.43, 0.36, 0.21)$$

b. Compute $p(y|x_1 = 0)$

$$p(y|x_1 = 0) = \frac{\boldsymbol{\pi} \circ \boldsymbol{\theta}}{\sum_j \pi_j \theta_j} = \frac{(0.25, 0.125, 0.125)}{0.5} = (0.5, 0.25, 0.25)$$

c. Compute $p(y|x_2 = 0)$

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = (0.24, 0.40, 0.24)$$

$$p(y|x_2 = 0) = \frac{\boldsymbol{\pi} \circ N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}{\sum_j \pi_j N(x_j|\mu_j, \sigma_j^2)} = (0.43, 0.36, 0.21)$$

d. Explain any interesting patterns you see in your results. Hint: look at the parameter vector $\boldsymbol{\theta}$

Since x_1 is independent of y we can see that it doesn't increase the amount of information we have about y . This shows in the $\boldsymbol{\theta}$ vector since (0, 1) are equally likely for any class. You always have a 50% chance of getting either a 0 or 1 regardless of the class.

2. 7.2 Multi-output linear regression

Solution:

$$\hat{W} = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y$$

$$\Phi(x) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\Phi(x)^T \Phi(x) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$(\Phi(x)^T \Phi(x))^{-1} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

$$\hat{\mathbf{W}} = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{bmatrix}$$

3. 7.7 Sufficient statistics for online linear regression

$$\bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y}^{(n)} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$C_{xx}^{(n)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad C_{xy}^{(n)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad C_{yy}^{(n)} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Solution:

(a) What are the minimal set of statistics that we need to estimate w_0 ?

$C_{xx}^{(n)}$ and $C_{xy}^{(n)}$ are sufficient statistics for w_0

(b) What are the minimal set of statistics that we need to estimate w_1 ?

$\bar{x}^{(n)}$, $\bar{y}^{(n)}$, $C_{xx}^{(n)}$ and $C_{xy}^{(n)}$ are sufficient statistics w_1

(c) Suppose a new data point, x_{n+1}, y_{n+1} arrives and we want to update our sufficient statistics without looking at the old data, which we have not stored. Show that we can do this for \bar{x} as follows.

$$\begin{aligned} \bar{x}^{(n+1)} &= \frac{1}{n+1} \sum_{i=1}^{n+1} x_i \\ &= \frac{1}{n+1} (n\bar{x}^{(n)} + x_{n+1}) && \text{the new sample summed with the current average current average then divided by the new total of samples. This gives equal weight to each sample} \\ &= \bar{x}^{(n)} + \frac{1}{n+1} (x_{n+1} - \bar{x}^{(n)}) \\ \bar{y}^{(n+1)} &= \frac{1}{n+1} \sum_{i=1}^{n+1} y_i \\ &= \frac{1}{n+1} (n\bar{y}^{(n)} + y_{n+1}) \\ &= \bar{y}^{(n)} + \frac{1}{n+1} (y_{n+1} - \bar{y}^{(n)}) \end{aligned}$$

This has the form: new estimate is old estimate plus correction. We see that the size of the correction diminishes over time (i.e., as we get more samples). Derive a similar expression to update \bar{y}

(d) Show that one can update C_{xy}^{n+1} recursively using

$$C_{xy}^{(n+1)} = \frac{1}{n+1} [x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{n+1}\bar{y}^{n+1}]$$

Let $n = n + 1$:

$$\begin{aligned} C_{xy}^{(n+1)} &= \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i - \bar{x}^{(n+1)})(y_i - \bar{y}^{(n+1)}) \\ &= \frac{1}{n+1} \left(\sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} \right) \\ &= \frac{1}{n+1} \left(x_{(n+1)}y_{(n+1)} + \sum_{i=1}^n x_i y_i - n\bar{x}^{(n)}\bar{y}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} \right) \\ &= \frac{1}{n+1} \left(x_{(n+1)}y_{(n+1)} + C_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} \right) \end{aligned}$$

Derive a similar expression to update C_{xx}

$$\begin{aligned} C_{xx}^{(n+1)} &= \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i - \bar{x}^{(n+1)})^2 \\ &= \frac{1}{n+1} \left(\sum_{i=1}^{n+1} x_i x_i - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)} \right) \\ &= \frac{1}{n+1} \left(x_{(n+1)}x_{(n+1)} + \sum_{i=1}^n x_i^2 - n\bar{x}^{(n)}\bar{x}^{(n)} + n\bar{x}^{(n)}\bar{x}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)} \right) \\ &= \frac{1}{n+1} \left(x_{(n+1)}x_{(n+1)} + C_{xx}^{(n)} + n\bar{x}^{(n)}\bar{x}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)} \right) \end{aligned}$$

4. Additional Question 1.

Compute the numbers of parameters that need to be estimated for the models listed below.

Solution:

(a) Naive Bayes (categorical input variables)

$pc + c - 1$ parameters are needed for this model

$$MLE = \theta_{jc}\pi_c$$

We also know that $\sum \pi_c = 1$

Since we don't know that about the θ value as the theta values do not necessarily sum to one because

they are dependent on the class that they are associated with, we know that we need $pc + c - 1$ parameters for this model

(b) Quadratic Discriminant Analysis (QDA)

$$\left(\frac{p(p+1)}{2} + p\right)c \text{ parameters are needed for this model}$$

We need the covariance matrix for sigma which is symmetric so we need the variance values and the half of the covariance values (since the matrix is symmetric we only need half the covariance values) then we need μ . We need these values for each class, and hence the above conclusion.

(c) Linear Discriminant Analysis (LDA)

$$\frac{p(p+1)}{2} + pc \text{ parameters are needed for this model}$$

We need the same covariance matrix for this model, but we assume the Σ parameter is the same for all the classes so we only need one covariance $p \times p$ matrix. We still need

(d) Diagonalized QDA (e.g. $\text{covariance}(X_i, X_j) = 0$ given $i \neq j$)

$$(p + p)c \text{ parameters are needed for this model}$$

Again, a more limited number of parameters than the QDA since we only need the diagonal of the covariance matrix for the parameters. Again we still need a μ value for each feature.

(e) Diagonalized LDA (e.g. $\text{covariance}(X_i, X_j) = 0$ given $i \neq j$)

$$p + pc \text{ parameters are needed for this model}$$

Again, we share the sigma so we need a weight for each feature and a μ for each class

(f) Linear Regression

$$p + 1 \text{ parameters are needed for this model}$$

Which is the size of β and this is equal to the number of features, the +1 is for β_0

(g) Ridge Regression

$$p + 2 \text{ parameters are needed for this problem}$$

p is the size of β and this is equal to the number of features. One +1 is for the β_0 parameter and the other +1 is the parameter estimating λ

5. Additional Question 2.

(bonus) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau \mathbf{I})$ where \mathbf{I} is the identity matrix, and Gaussian sampling model $y \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula and the variances τ and σ^2 .

Solution:

6. Additional Question 3.

Solution:

(a) Derive the OLS linear regression estimate.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^N -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^N -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^N -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^N y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

$$N\hat{\beta}_0 = N\bar{y} - N\hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^N -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^N x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^N x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^N x_i y_i - \bar{y} \sum_{i=1}^N x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^N x_i - \hat{\beta}_1 \sum_{i=1}^N x_i^2 = 0$$

$$\sum_{i=1}^N y_i = N\bar{y}$$

$$\sum_{i=1}^N x_i = N\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

(b) Derive the Ridge regression estimate.

$$\begin{aligned} RSS &= \|y - \hat{y}\|_2^2 + \lambda \|\beta\|_2^2 \\ &= (Y - X\beta)^T (Y - X\beta) + \lambda (\beta)^T \beta \\ &= (Y - X\beta)^T (Y - X\beta) + \lambda (\beta)^T \beta \\ &= (Y^T - \beta^T X^T)(Y - X\beta) + \lambda (\beta^T \beta) \\ &= Y^T Y - \beta^T X^T Y + Y^T X \beta + \beta^T X^T X \beta + \lambda (\beta^T \beta) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta^T} &= 0 - X^T Y + 0 + X^T X \beta + \lambda \beta = 0 \\ &= X^T Y + X^T X \beta + \lambda \beta = 0 \end{aligned}$$

$$\begin{aligned} X^T Y &= X^T X \beta + \lambda \beta \\ &= \beta (X^T X + \lambda I) \end{aligned}$$

$$\hat{\beta} = (X^T X + \lambda I)^{-1} (X^T Y)$$

(c) Show that the estimation of β_λ^{ridge} is a biased estimator. [Hint: compute the expected value of the ridge estimate and show that it is not equal to the expected value of the OLS estimate.]

$$\begin{aligned} \hat{\beta}_\lambda^{ridge} &= (X^T X + \lambda I)^{-1} (X^T y) \\ &= (X^T X + \lambda I_p)^{-1} (X^T X) (X^T X)^{-1} X^T y \\ &= (X^T X (I + (X^T X)^{-1} \lambda))^{-1} (X^T X) \hat{\beta}^{OLS} \\ &= (I + (X^T X)^{-1} \lambda)^{-1} \hat{\beta}^{OLS} \end{aligned}$$

$$(I + (X^T X)^{-1} \lambda)^{-1} \hat{\beta}^{OLS} \neq \hat{\beta}^{OLS} \quad \text{unless } \lambda = 0$$