Assignment 1

Adriana Salcedo

November 11, 2017

1 Q1

1.1

- y is a response for a single observation
- $\mathbf{x}$ is an N x d vector, in the case of a single observation (below) it is a 1 x d vector
- $\boldsymbol{\mu}_k$ is a 1 x d vector of the means of each feature for class k
- $\boldsymbol{\sigma}$ is a d x 1 vector of the variances of each feature
- $\boldsymbol{\alpha}_k$ is a d x 1 vector of the prior for each class

$$p(y=k|\mathbf{x},\boldsymbol{\mu}_k,\boldsymbol{\sigma}) = \frac{p(\mathbf{x}|y=k,\boldsymbol{\mu}_k,\boldsymbol{\sigma})p(y=k)}{p(\mathbf{x})}$$

$$= \frac{\mathbb{1}\{y_k=k\}(2\pi\boldsymbol{\sigma}^T\boldsymbol{\sigma})^{-\frac{1}{2}}exp\{-\frac{1}{2\boldsymbol{\sigma}^T\boldsymbol{\sigma}}(\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^T\}\boldsymbol{\alpha}_k}{\sum_{k=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\sigma})\boldsymbol{\alpha}_k}$$

$$= \frac{\mathbb{1}\{y_k=k\}(2\pi\boldsymbol{\sigma}^T\boldsymbol{\sigma})^{-\frac{1}{2}}exp\{-\frac{1}{2\boldsymbol{\sigma}^T\boldsymbol{\sigma}}(\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^T\}\boldsymbol{\alpha}_k}{\sum_{k=1}^K(2\pi\boldsymbol{\sigma}^T\boldsymbol{\sigma})^{-\frac{1}{2}}exp\{-\frac{1}{2\boldsymbol{\sigma}^T\boldsymbol{\sigma}}(\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^T\}\boldsymbol{\alpha}_k}$$

1.2

For each class k

$$\ell(\boldsymbol{\theta},D) = \prod_{n=1}^N \mathbb{1}\{t_k=k\}p(\mathbf{x}_n|y_n=k,\boldsymbol{\mu}_k,\boldsymbol{\sigma})p(y=k)$$

$$\prod_{n=1}^N \mathbb{1}\{t_k=k\}(2\pi\boldsymbol{\sigma}^T\boldsymbol{\sigma})^{-\frac{1}{2}}exp\{-\frac{1}{2\boldsymbol{\sigma}^T\boldsymbol{\sigma}}(\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^T\}\boldsymbol{\alpha}_k$$

Apply negative log

$$= -\sum_n^N \mathbb{1}\{y_k=k\}\log p(\mathbf{x}_n|y_n=k,\boldsymbol{\mu}_\mathbf{k},\boldsymbol{\sigma}) - \mathbb{1}\{y_k=k\}\log\alpha_k$$

$$= -\sum_n^N \mathbb{1}\{y_k = k\} \log (2\pi \boldsymbol{\sigma}^T \boldsymbol{\sigma})^{-\frac{1}{2}} + \mathbb{1}\{y_k = k\} \frac{1}{2\boldsymbol{\sigma}^T \boldsymbol{\sigma}} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T - \mathbb{1}\{y_k = k\} \log \boldsymbol{\alpha}_k$$

$$= \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \log (2\pi \boldsymbol{\sigma}^T \boldsymbol{\sigma}) + \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2\boldsymbol{\sigma}^T \boldsymbol{\sigma}} (\mathbf{x}_n \mathbf{x}_n^T - 2\mathbf{x}_n \boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T) - \mathbb{1}\{y_k = k\} \log \boldsymbol{\alpha}_k$$

Taking the derivative with respect to μ_{ki}

$$\sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2\boldsymbol{\sigma}^T \boldsymbol{\sigma}} (-2x_{ki} + 2\mu_{ki})$$

Taking the derivative with respect to σ_{ki}

$$= \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \frac{1}{(2\pi \boldsymbol{\sigma}^T \boldsymbol{\sigma})} 2\pi - \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} (\boldsymbol{\sigma}^T \boldsymbol{\sigma})^{-2} (\mathbf{x}_n \mathbf{x}_n^T - 2\mathbf{x}_n \boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T)$$

$$= \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \frac{1}{(\boldsymbol{\sigma}^T \boldsymbol{\sigma})} - \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} (\boldsymbol{\sigma}^T \boldsymbol{\sigma})^{-2} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T$$

1.3

Setting the derivative with respect to μ_{ki} to zero

$$0 = \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2\boldsymbol{\sigma}^T \boldsymbol{\sigma}} 2x_{ki} + \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2\boldsymbol{\sigma}^T \boldsymbol{\sigma}} 2\mu_{ki}$$

$$\sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{\boldsymbol{\sigma}^T \boldsymbol{\sigma}} \mu_{ki} = \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{\boldsymbol{\sigma}^T \boldsymbol{\sigma}} x_{ki}$$

$$\mu_{ki} = \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{ki}}{\sum_n^N \mathbb{1}\{y_k = k\}}$$

For for feature i=1 to i=d, and for each class k =1 to k = k

$$\boldsymbol{\mu}_k = \begin{bmatrix} \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=1,i=1}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=1,i=2}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \dots & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=1,i=d}}{\sum_n^N \mathbb{1}\{y_k = k\}} \\ \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=2,i=1}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=2,i=2}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \dots & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=2,i=d}}{\sum_n^N \mathbb{1}\{y_k = k\}} \\ \dots & & & \\ \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=k,i=1}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=k,i=2}}{\sum_n^N \mathbb{1}\{y_k = k\}} & \dots & \frac{\sum_n^N \mathbb{1}\{y_k = k\} x_{k=k,i=d}}{\sum_n^N \mathbb{1}\{y_k = k\}} \end{bmatrix}$$

Setting the derivative with respect to σ_i^2 to zero

$$0 = \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \frac{1}{(\boldsymbol{\sigma}^T \boldsymbol{\sigma})} - \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} (\boldsymbol{\sigma}^T \boldsymbol{\sigma})^{-2} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T$$

$$\sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \frac{1}{(\boldsymbol{\sigma}^T \boldsymbol{\sigma})} = \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{2} \frac{1}{\boldsymbol{\sigma}^T \boldsymbol{\sigma}^2} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T$$

$$\sum_n^N \mathbb{1}\{y_k = k\} = \sum_n^N \mathbb{1}\{y_k = k\} \frac{1}{\boldsymbol{\sigma}^T \boldsymbol{\sigma}} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T$$

$$\boldsymbol{\sigma}^T \boldsymbol{\sigma} = \frac{\sum_n^N \mathbb{1}\{y_k = k\} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T}{\sum_n^N \mathbb{1}\{y_k = k\}}$$

For each feature i

$$\sigma_i = \frac{\sum_n^N \mathbb{1}\{y_k = k\} (x_i - \mu_{ki})^2}{\sum_n^N \mathbb{1}\{y_k = k\}}$$

2

2.1.1

K=1 train accuracy: 0.9998, test accuracy: 0.9685
K=15 train accuracy: 0.9596 test accuracy: 0.958

2.1.2

I chose to break ties by randomly selecting a class from among the tied classes. I chose this approach as it gave each of the each of the classes tied for the most votes an equal chance of being selected, allowed the algorithm to use the same k for every point, did not exclude any test points, and always gave a class for each test point (ie there were no cases where a class was undefined).

2.1.3

The optimal k was 2. train accuracy: 0.9817, test accuracy: 0.9617

2.2

MLE estimates for μ_k and Σ_k

$\mathbf{x}_i$ is a d x 1 vector of the features for one observation and μ_{ki} is a d x 1 vector of the means of the features for class k. y is a N x 1 vector of responses for each observation indicating the class

$$\ell(\mathbf{x}|\boldsymbol{\mu}, \sigma_k) = p(\mathbf{x}|y = k, \boldsymbol{\mu}_k, \Sigma) p(y = k)$$

Applying the log

$$= \log p(\mathbf{x}|y = k, \boldsymbol{\mu}, \sigma) + \log p(y = k)$$

$$= \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \log{(2\pi)^{-\frac{d}{2}}} |\Sigma_k|^{-\frac{1}{2}} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + \log \frac{1}{10}$$

$$= - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \log{(2\pi)^{\frac{d}{2}}} |\Sigma_k|^{\frac{1}{2}} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + \log \frac{1}{10}$$

taking the derivative w.r.t μ_k and setting it to zero

$$0 = - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \Sigma_k^{-1} \frac{1}{2} (-2\mathbf{x}_i + 2\boldsymbol{\mu}_k)$$

$$0 = \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \Sigma_k^{-1} \mathbf{x}_i - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \Sigma_k^{-1} \boldsymbol{\mu}_k$$

$$\sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \Sigma_k^{-1} \mathbf{x}_i = \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \Sigma_k^{-1} \boldsymbol{\mu}_k$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \mathbf{x}_i}{\sum_{i=1}^{N} \mathbb{1}\{y_k = k\}}$$

To solve for the covariance

$$- \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \log{(2\pi)^{\frac{d}{2}}} |\Sigma_k|^{\frac{1}{2}} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (x_i - \mu_{ki})^T \Sigma_k^{-1} (x_i - \mu_{ki}) + \log \frac{1}{10}$$

taking the derivative w.r.t Σ_{ki}^{-1} and setting it to zero

$$0 = - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{\partial \log{(2\pi)^{\frac{d}{2}}} |\Sigma_k|^{\frac{1}{2}}}{\partial \Sigma_k^{-1}} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

$$0 = \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} (2\pi)^{-\frac{d}{2}} |\Sigma_k|^{-\frac{1}{2}} (2\pi)^{\frac{d}{2}} \frac{\partial |\Sigma_k^{-1}|^{-\frac{1}{2}}}{\partial \Sigma_k^{-1}} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

$$0 = \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} |\Sigma_k^{-1}|^{\frac{1}{2}} (-\frac{1}{2}) |\Sigma_k^{-1}|^{-\frac{3}{2}} |\Sigma_k^{-1}| \Sigma_k^{T} - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

$$0 = \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} (-\frac{1}{2}) \Sigma_k - \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

$$\Sigma_k = \frac{\mathbb{1}\{y_k = k\} \sum_{i=1}^{N} \mathbb{1}\{y_k = k\} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \mathbb{1}\{y_k = k\}}$$

2.2.3