Statistical Methods: Assessed Coursework 2

By Henry Bourne

Question 0

E

Question 1

1.1

The answer is c or d as this line goes roughly through the class means and when the data points are projected onto the embedding vector (for either of the directions) it appears as though the between-class scatterness will be maximized and the within-class scatterness minimized as points from the negative/positive class will be close together and the class means in the embedding will be fairly centrally located (within their classes). For comparison we can argue that the opposite would be true for directions a and b, where points from both classes will not be separated (class-wise) in the embedding and the furthest points from a given class to its class mean will be larger than we had with embedding vectors in directions c or d.

1.2

We can construct a likelihood function over the entire dataset:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \prod_{i=1}^{n} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} N_{\boldsymbol{x}_i}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

ie. all the points are iid. normally distributed (have same mean and covariance matrix) where $\boldsymbol{\Sigma}_{ML}$ maximizes the above likelihood.

Let's take the decomposition to be the eigen-decomposition of $\boldsymbol{\Sigma}_{ML}$, then $\boldsymbol{\mu}_1$ is an eigenvector of $\boldsymbol{\Sigma}_{ML}$ with corresponding eigenvalue D_1. Note that $D_1 > D_2$ which means the eigenvector $\boldsymbol{u}_1$ corresponds to the direction of the largest variance in the data, hence must correspond to direction a or b.

Question 3

Recall the soft margin SVM:

$$\text{Minimize } ||\boldsymbol{w}'||^2 + \sum_{i \in D} \epsilon_i \tag{2}$$

$$\text{Subject to } \forall i, y_i \cdot f(\boldsymbol{x}_i; \boldsymbol{w}) + \epsilon_i \geq 1, \epsilon_i \geq 0 \tag{3}$$

We can modify the objective function such that it penalizes False Negatives (FN) by making FNs 1000 times more costly by rewriting the objective function as:

$$\text{Minimize } ||\boldsymbol{w}'||^2 + \sum_{i \in D} (1000 \cdot I(y_i = 1) + (1 - I(y_i = 1))) \cdot \epsilon_i \tag{4}$$

where $I(y_i = 1)$ is the indicator function that is one when $y_i = 1$ (ie. when wrongly classifying would lead to a FN) it multiplies the cost times 1000. So, during optimization we will have attributed a 1000x cost to giving a FN, meanwhile the cost of a FP stays as is (a 1000th of the cost).

Question 4

4.1

B

4.2

The factorization of $p(y, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)})$ according the to graph is:

$$p(y) \cdot p(x^{(1)}|y) \cdot p(x^{(2)}|y, x^{(1)}) \cdot p(x^{(3)}|x^{(2)}) \cdot p(x^{(4)}|x^{(1)}) \tag{5}$$

The conditional independencies encoded by this graph are:

$$x^{(2)} \perp x^{(4)}|y, x^{(1)} \tag{6}$$

$$x^{(3)} \perp y, x^{(1)}, x^{(4)}|x^{(2)} \tag{7}$$

$$x^{(4)} \perp y, x^{(2)}, x^{(3)}|x^{(1)} \tag{8}$$

And no we should just use $x^{(1)}$ and $x^{(2)}$ as given these two features we have from our conditional independencies that $x^{(3)}$ and $x^{(4)}$ are independent of y, we also have that our prediction function:

$$p(y|x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) = \frac{p(y) \cdot p(x^{(1)}|y) \cdot p(x^{(2)}|y, x^{(1)}) \cdot p(x^{(3)}|x^{(2)}) \cdot p(x^{(4)}|x^{(1)})}{p(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)})} \tag{9}$$

$$\propto p(y) \cdot p(x^{(1)}|y) \cdot p(x^{(2)}|y, x^{(1)}) \tag{10}$$

Therefore to formulate a prediction for y we only need $x^{(1)}$ and $x^{(2)}$.