Homework 4 - Group 076

Aprendizagem 2021/2022

1 Pen and Paper

1) Let $\mathbf{c}$ be a 2-dimensional random vector for which $c_k \in \{0, 1\}$ is an indicator of cluster k ($k \in \{1, 2\}$). Let $\pi_k = p(c_k = 1)$ and let $\gamma(c_{nk}) = p(c_k = 1|\mathbf{x}^{(n)})$.

- **E-step**

  In this step, we assign each instance $\mathbf{x}^{(n)}$ to the cluster c_k that yields the largest posterior $\gamma(c_{nk})$. Given that $\mathbf{x}^{(n)}|c_k = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$, by Bayes rule, we have:

$$\gamma(c_{nk}) = p(c_k = 1|\mathbf{x}^{(n)}) = \frac{p(c_k = 1)p(\mathbf{x}^{(n)}|c_k = 1)}{\sum_{l=1}^{2} p(c_l = 1)p(\mathbf{x}^{(n)}|c_l = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k)}{\sum_{l=1}^{2} \pi_l \mathcal{N}(\mathbf{x}^{(n)}|\mu_l, \Sigma_l)}$$

  Setting $\mu_k = \mathbf{x}^{(k)}$ ($k \in \{1, 2\}$), we now compute the posteriors:

  – For $\mathbf{x}^{(1)}$:

$$\gamma(c_{11}) = \frac{\pi_1 \mathcal{N}(\mathbf{x}^{(1)}|\mu_1, \Sigma_1)}{\sum_{l=1}^{2} \pi_l \mathcal{N}(\mathbf{x}^{(1)}|\mu_l, \Sigma_l)} = \frac{0.7 \cdot 0.15915}{0.7 \cdot 0.15915 + 0.3 \cdot 0.0} = \frac{0.11141}{0.11141} = 1.0$$
$$\gamma(c_{12}) = 1 - \gamma(c_{11}) = 0.0$$

  – For $\mathbf{x}^{(2)}$:

$$\gamma(c_{21}) = \frac{0.7 \cdot 0.0}{0.7 \cdot 0.0 + 0.3 \cdot 0.07958} = 0.0 \quad \gamma(c_{22}) = 1 - \gamma(c_{21}) = 1.0$$

  – For $\mathbf{x}^{(3)}$:

$$\gamma(c_{31}) = \frac{0.7 \cdot 0.00024}{0.7 \cdot 0.00024 + 0.3 \cdot 0.00001} = 0.98271 \quad \gamma(c_{32}) = 1 - \gamma(c_{31}) = 0.01729$$

  – For $\mathbf{x}^{(4)}$:

$$\gamma(c_{41}) = \frac{0.7 \cdot 0.00001}{0.7 \cdot 0.00001 + 0.3 \cdot 0.0} = 0.85698 \quad \gamma(c_{42}) = 1 - \gamma(c_{41}) = 0.14302$$

- **M-step**

  We re-compute μ_k, Σ_k and π_k ($k \in \{1, 2\}$) so as to increase the likelihood of the data. The new cluster centroids are given by:

$$\mu_k = \frac{\sum_{n=1}^{4} \gamma(c_{nk})\mathbf{x}^{(n)}}{\sum_{n=1}^{4} \gamma(c_{nk})}$$

  And so we have:

$$\mu_1 = \frac{1.0 \begin{bmatrix} 2 \\ 4 \end{bmatrix} + 0.0 \begin{bmatrix} -1 \\ -4 \end{bmatrix} + 0.98271 \begin{bmatrix} -1 \\ 2 \end{bmatrix} + 0.85698 \begin{bmatrix} 4 \\ 0 \end{bmatrix}}{1.0 + 0.0 + 0.98271 + 0.85698} = \begin{bmatrix} 1.56538 \\ 2.10073 \end{bmatrix}$$

$$\mu_2 = \frac{0.0 \begin{bmatrix} 2 \\ 4 \end{bmatrix} + 1.0 \begin{bmatrix} -1 \\ -4 \end{bmatrix} + 0.01729 \begin{bmatrix} -1 \\ 2 \end{bmatrix} + 0.14302 \begin{bmatrix} 4 \\ 0 \end{bmatrix}}{0.0 + 1.0 + 0.01729 + 0.14302} = \begin{bmatrix} -0.3837 \\ -3.41758 \end{bmatrix}$$

The new covariance matrices are now given by:

$$\Sigma_k = \frac{\sum_{n=1}^{4} \gamma(c_{nk})(\mathbf{x}^{(n)} - \mu_k)(\mathbf{x}^{(n)} - \mu_k)^T}{\sum_{n=1}^{4} \gamma(c_{nk})}$$

And so:

$$\Sigma_1 = \frac{1.0 \begin{bmatrix} 2.0 & 8.0 \\ 8.0 & 16.0 \end{bmatrix} + 0.0 \begin{bmatrix} 1.0 & 4.0 \\ 4.0 & 16.0 \end{bmatrix} + 0.98271 \begin{bmatrix} 1.0 & -2.0 \\ -2.0 & 4.0 \end{bmatrix} + 0.85698 \begin{bmatrix} 4.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}}{1.0 + 0.0 + 0.98271 + 0.85698} = \begin{bmatrix} 4.13282 & -1.16337 \\ -1.16337 & 2.6056 \end{bmatrix}$$

$$\Sigma_2 = \frac{0.0 \begin{bmatrix} 2.0 & 8.0 \\ 8.0 & 16.0 \end{bmatrix} + 1.0 \begin{bmatrix} 1.0 & 4.0 \\ 4.0 & 16.0 \end{bmatrix} + 0.01729 \begin{bmatrix} 1.0 & -2.0 \\ -2.0 & 4.0 \end{bmatrix} + 0.14302 \begin{bmatrix} 4.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}}{0.0 + 1.0 + 0.01729 + 0.14302} = \begin{bmatrix} 2.70166 & 2.10624 \\ 2.10624 & 2.16924 \end{bmatrix}$$

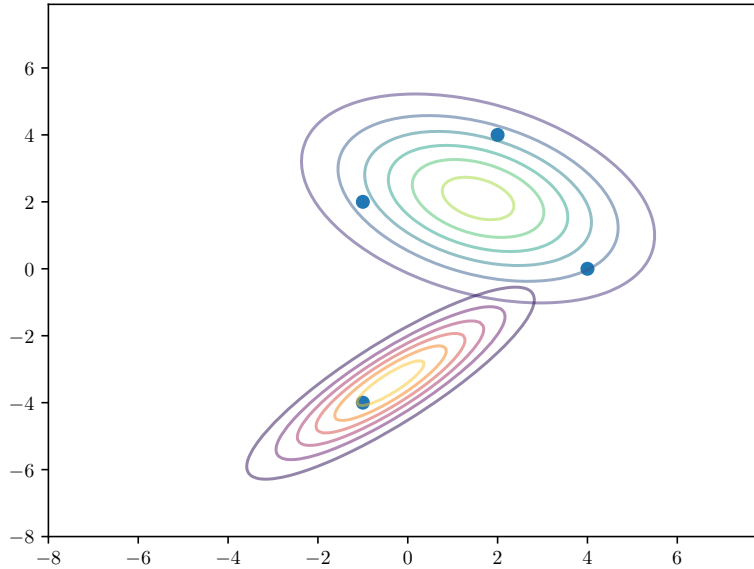Finally, the new prior probabilities, can now be written as:

$$\pi_k = \frac{\sum_{n=1}^{4} \gamma(c_{nk})}{4}$$

and thus:

$$\pi_1 = \frac{1.0 + 0.0 + 0.98271 + 0.85698}{4} = 0.70992$$

$$\pi_2 = \frac{0.0 + 1.0 + 0.01729 + 0.14302}{4} = 0.29008$$

Hence, this iteration of the EM algorithm assigned points $\mathbf{x}^{(1)}$, $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$ to cluster c_1, while $\mathbf{x}^{(2)}$ was assigned to cluster c_2. The following figure depicts this clustering:

2) First, we compute the silhouette of cluster c_1:

$$s(\mathbf{x}^{(1)}) = 1 - \frac{a(\mathbf{x}^{(1)})}{b(\mathbf{x}^{(1)})} = 1 - \frac{\frac{\|\mathbf{x}^{(1)}-\mathbf{x}^{(3)}\|_2 + \|\mathbf{x}^{(1)}-\mathbf{x}^{(4)}\|_2}{2}}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2} = 1 - \frac{\frac{3.60555+4.47214}{2}}{8.544} = 0.52729$$

$$s(\mathbf{x}^{(3)}) = 1 - \frac{a(\mathbf{x}^{(3)})}{b(\mathbf{x}^{(3)})} = 1 - \frac{\frac{\|\mathbf{x}^{(3)}-\mathbf{x}^{(1)}\|_2 + \|\mathbf{x}^{(3)}-\mathbf{x}^{(4)}\|_2}{2}}{\|\mathbf{x}^{(3)} - \mathbf{x}^{(2)}\|_2} = 1 - \frac{\frac{3.60555+5.38516}{2}}{4.49536} = 0.25077$$

$$s(\mathbf{x}^{(4)}) = 1 - \frac{a(\mathbf{x}^{(4)})}{b(\mathbf{x}^{(4)})} = 1 - \frac{\frac{\|\mathbf{x}^{(4)}-\mathbf{x}^{(1)}\|_2 + \|\mathbf{x}^{(4)}-\mathbf{x}^{(3)}\|_2}{2}}{\|\mathbf{x}^{(4)} - \mathbf{x}^{(2)}\|_2} = 1 - \frac{\frac{4.47214+5.38516}{2}}{6.40312} = 0.23027$$

$$s(c_1) = \frac{s(\mathbf{x}^{(1)}) + s(\mathbf{x}^{(3)}) + s(\mathbf{x}^{(4)})}{3} = \frac{0.52729 + 0.25077 + 0.23027}{3} = 0.33611$$

Since $|c_2| = 1$, we set $s(c_2) = 1$. The silhouette of the solution is defined as the arithmetic mean of the silhouettes of the clusters:

$$s(C) = \frac{s(c_1) + s(c_2)}{2} = \frac{0.33611 + 1.0}{2} = 0.66806$$

3) (a) i) The specified MLP is described by the following parameters: weight matrices $W^{[1]}$, $W^{[2]}$, $W^{[3]}$ and $W^{[4]}$, along bias vectors $b^{[1]}$, $b^{[2]}$, $b^{[3]}$ and $b^{[4]}$. For input data of dimention d, we have:
- $W^{[1]}, W^{[2]}, W^{[3]} \in \mathbb{R}^{d \times d}$
- $W^{[4]} \in \mathbb{R}^{2 \times d}$
- $b^{[1]}, b^{[2]}, b^{[3]} \in \mathbb{R}^d$
- $b^{[4]} \in \mathbb{R}^2$

Consequently, the total number of parameters is $3d^2 + 2d + 3d + 2 = 3d^2 + 5d + 2$. Given that the number of parameters is a reasonable estimate of the VC dimension of the model, for $d = 5$, we have an estimate of the VC dimension of the aforementioned MLP of $3 \cdot 5^2 + 5 \cdot 5 + 2 = 102$

ii) For input data of dimensionality d, we know that $d_{\mathrm{VC}} \leq 3^d$, as 3^d is precisely the size of the input space (there are d features with 3 possible values). Hence, it is impossible for the decision tree to shatter a greater number of points than 3^d. On another hand, we note that any set of 3^d can be shattered by tree; for that purpose, given any dichotomy, it suffices to consider a tree with $d+1$ levels with a single non-leaf node which is associated with a point in the set. In that node, we test equality w.r.t. that point, and we classify the example with the label of the corresponding point or continue down the tree accordingly. Thus, $d_{\mathrm{VC}} = 3^d = 3^5 = 243$.

iii) We again estimate VC dimension based on the number of parameters of the model. Given a point $\mathbf{x} \in \mathbb{R}^d$ and a binary output variable $z \in \{0, 1\}$, the Bayesian Classifier estimates:

$$p(z = 1|\mathbf{x}) = \frac{p(\mathbf{x}|z = 1)p(z = 1)}{p(z = 0)p(\mathbf{x}|z = 0) + p(z = 1)p(\mathbf{x}|z = 1)}$$

Thus, we need to estimate the prior $p(z)$ and the likelihood $p(\mathbf{x}|z)$. Since $z \sim \text{Bernoulli}(p)$, there is only one parameter associated with the prior; we also have that $\mathbf{x}|z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$ ($k \in \{0, 1\}$), where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$. Since Σ_k is symmetric, we can discard the lower diagonal part of the matrix. Thus, there are $1 + 2(d + d + \frac{d^2-d}{2}) = 1 + 4d + d^2 - d = d^2 + 3d + 1$ parameters, and we have an approximate VC dimension of $5^2 + 3 \cdot 5 + 1 = 41$.

(b) In comparison the other models, the VC dimension of the decision tree starts to explode for data dimensionalities as low as 10, requiring exponentially larger datasets to mitigate overfitting risks.
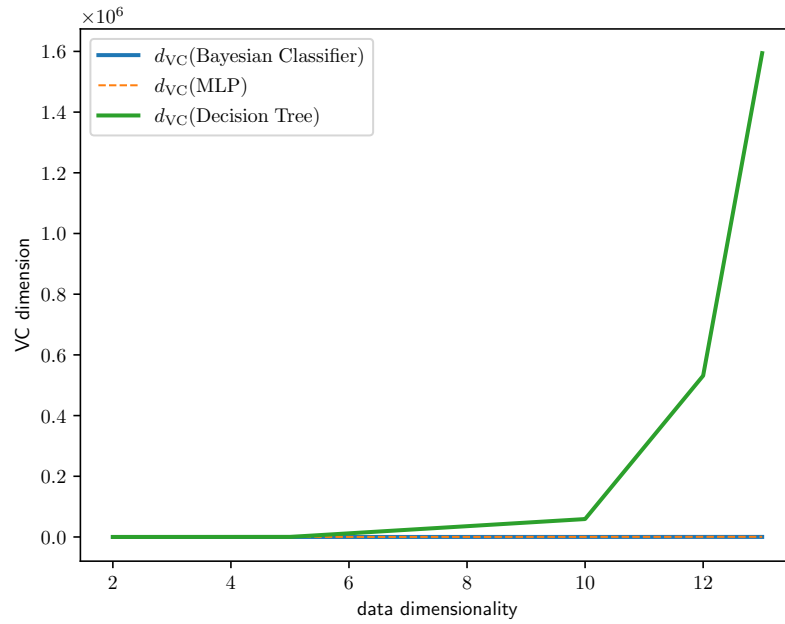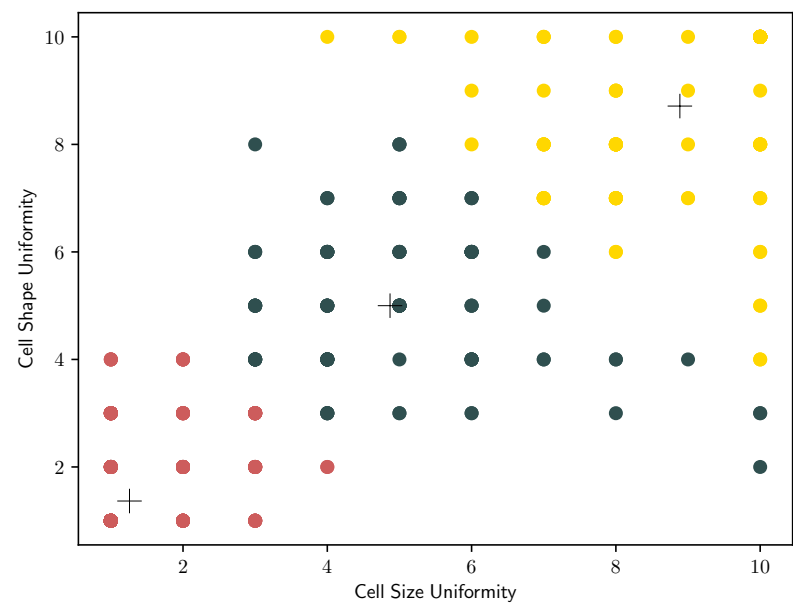
(c) From the plot below, we can that the VC dimension of the decision tree starts
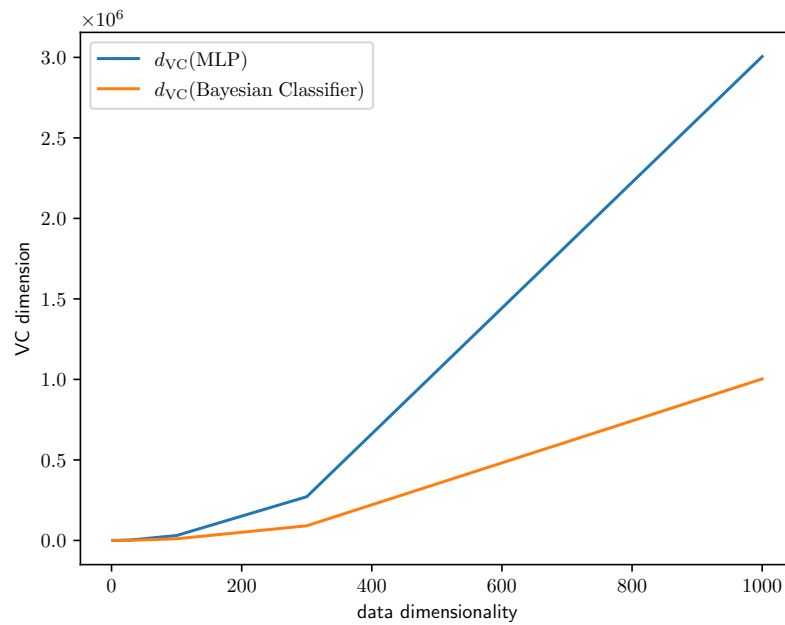
2 Programming and critical analysis

4) (a) The produced solution yielded an ECR value of 13.5 for 2-means and 6.67 for 3-means.

(b) The produced solution yielded a Silhouette value of 0.5968 for 2-means and 0.52454 for 3-means.

5)

6) Lezz go