1 Bayes Classifier

In the lecture we saw that the Bayes classifier is

$$c^*(x) := \arg\max_{y \in \mathcal{Y}} p(y|x).$$ \hspace{1cm} (1)

a) Which of these decision functions is equivalent to $c^*$?

- $c_1(x) := \arg\max_y p(x)$
- $c_2(x) := \arg\max_y p(y)$
- $c_3(x) := \arg\max_y p(x, y)$
- $c_4(x) := \arg\max_y p(x | y)$

For $\mathcal{Y} = \{-1, +1\}$, we can express the Bayes classifier as $c^*(x) = \text{sign}\left[\log p(+1|x) - \log p(-1|x)\right]$

b) Which of the following expressions are equivalent to $c^*$?

- $c_5(x) := \text{sign}\left[\log p(x, +1) - \log p(x, -1)\right]$
- $c_6(x) := \text{sign}\left[\log p(+1|x) + \log p(-1|x)\right]$
- $c_7(x) := \text{sign}\left[\log p(+1|x) - \log p(-1|x)\right]$
- $c_8(x) := \text{sign}\left[\log p(x, +1) - \log p(x, -1)\right]$

2 Gaussian Discriminant Analysis

**Gaussian Discriminant Analysis (GDA)** is an easy-to-compute method for generative probabilistic classification. For a training set $\mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\} \subset \mathbb{R}^d \times \{1, \ldots, M\}$, set

$$\mu := \frac{1}{n} \sum_{i=1}^n x^i, \quad \Sigma := \frac{1}{n} \sum_{i=1}^n (x^i - \mu)(x^i - \mu)\top, \quad \mu_y := \frac{1}{|\{i : y^i = y\}|} \sum_{\{i : y^i = y\}} x^i, \quad \text{for } y \in \mathcal{Y},$$ \hspace{1cm} (2)

and define

$$p(x|y) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left(-\frac{1}{2} (x - \mu_y)\top \Sigma^{-1} (x - \mu_y)\right)$$ \hspace{1cm} (3)

a) Show for binary classification ($M = 2$): GDA leads to a linear decision rule, regardless of what $p(y)$ is.

b) GDA is popular when there are many classes but only few examples for each class. Can you imagine why?
3 Practical Experiments III

- Pick one more training methods from the previous sheet and implement it.
- Implement Gaussian Discriminant Analysis as in exercise 2.
- What error rates do both methods achieve on the datasets from the previous sheet?

4 Practical Experiments IV

- Create an ”XOR”-dataset in $\mathbb{R}^2$ (as in the figure on the right) that has:
  - 50 points of class 1 uniformly randomly located in $[0, 1] \times [0, 1]$
  - another 50 points of class 1 uniformly randomly located in $[-1, 0] \times [-1, 0]$
  - 50 points of class $-1$ uniformly randomly located in $[-1, 0] \times [0, 1]$
  - another 50 points of class $-1$ uniformly randomly located in $[0, 1] \times [-1, 0]$

- Split the dataset randomly into 2 parts: 50% for training, 50% as test set.

- Implement a Gaussian Mixture Model (GMM) with $k$ components in $\mathbb{R}^d$. For training, use the EM-algorithm as introduced in Lecture 2.

- For each $y \in \{\pm 1\}$, fit one GMM with $k = 2$ to the corresponding points of the XOR-datasets.

- Evaluate the classifier that is induced by the GMM. What is its error rate on the test data?

5 Optional: Uniform-Weight Gaussian Mixture Model

Imagine you want to learn a GMM, but all $k$ components should have the same mixture weights, $\pi = (\frac{1}{k}, \ldots, \frac{1}{k})$. What happens if you try to find the maximum likelihood solution by simply taking the derivative of the likelihood? What happens to the EM algorithm? Can you come up with a better algorithm?

6 Refresher: Convex Duality

Refresh your knowledge on convexity, Lagrangian multipliers and convex duality. You don’t have to hand in anything, but it’ll be a useful preparation for the next lecture and exercise sheet.