IST Austria: Statistical Machine Learning 2018/19

Christoph Lampert <chl@ist.ac.at>
TAs: Nikola Konstantinov <nkonstan@ist.ac.at>,
Mary Phuong <bphuong@ist.ac.at>
Amelie Royer <aroyer@ist.ac.at>

Exercise Sheet 2/5 (due date 22/10/2018)

1 Bayes Classifier

In the lecture we saw that the Bayes classifier is

$$c^*(x) := \operatorname{argmax}_{y\in\mathcal{Y}} p(y|x). \tag{1}$$

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

- $c_1(x) := \operatorname{argmax}_y p(x)$
- $c_2(x) := \operatorname{argmax}_y p(y)$
- $c_3(x) := \operatorname{argmax}_y p(x, y)$
- $c_4(x) := \operatorname{argmax}_y p(x|y)$

For $\mathcal{Y} = \{-1, +1\}$, we can express the Bayes classifier as $c^*(x) = \operatorname{sign}[\log \frac{p(+1|x)}{p(-1|x)}]$
b) Which of the following expressions are equivalent to c^*?

- $c_5(x) := \operatorname{sign}[\frac{\log p(x,+1)}{\log p(x,-1)}]$
- $c_6(x) := \operatorname{sign}[\log p(+1|x) + \log p(-1|x)]$
- $c_7(x) := \operatorname{sign}[\log p(+1|x) - \log p(-1|x)]$
- $c_8(x) := \operatorname{sign}[\log p(x, +1) - \log p(x, -1)]$
- $c_9(x) := \operatorname{sign}[p(+1|x) - p(-1|x)]$
- $c_{10}(x) := \operatorname{sign}[\frac{p(x,+1)}{p(x,-1)} - 1]$
- $c_{11}(x) := \operatorname{sign}[\frac{\log p(+1|x)}{\log p(-1|x)} - 1]$
- $c_{12}(x) := \operatorname{sign}[\log \frac{p(x|+1)}{p(x|-1)} + \log \frac{p(+1)}{p(-1)}]$

2 Gaussian Discriminant Analysis

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

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

and define

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

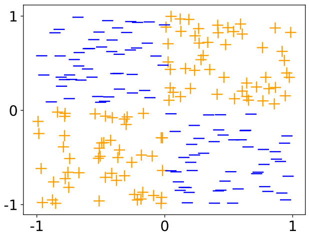
a) Show for binary classification ($M = 2$): GDA leads to a linear decision rule, regardless of what $p(y)$ is.
b) GDA is popular when there are many classes but only few examples for each class. Can you imagine why?

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, $\pi = (\frac{1}{k}, \dots, \frac{1}{k})$. 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.