Exercise Sheet #5

Due: 28.11.22 11h00

Roman Remme

General Regulations.

- Please hand in your solutions in groups of three people. A mix of attendees from Monday and Tuesday tutorials is fine.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/hci-unihd/mlph_sheet05`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in both the notebook (`.ipynb`), as well as an exported pdf.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of three.

1 QDA

In this exercise we apply QDA to a one dimensional binary classification problem.

(a) For the 1D binary classification problem given in `data1d.npy`, `labels1d.npy`, fit a normal distribution to each class by computing the mean and standard deviation of the points belonging to it. (2 pts)

(b) QDA: In the range $[-10, 10]$, compute and plot the two Gaussian class densities as well as the posterior $p(y = 0|x)$ assuming equal prior probabilities, i.e. $p(y = 0) = p(y = 1)$. What do you observe? (2 pts)

2 Mean of the Bernoulli distribution

Let X be a random Variable distributed as the Bernoulli distribution $P(X = x) = \mathrm{Bern}(x; \mu)$. What is the expectation of X? (2 pts)

3 Trees and Random Forests (10 pt)

(a) Calculate optimal splits: For the provided (`data1d.npy`, `labels1d.npy`) one-dimensional binary classification problem, consider all splits where the smallest $i = 1, \ldots, N-1$ data points are grouped into one node and the remaining $N - i$ points into the other. For each of these splits, compute the Gini impurity, entropy and misclassifcation rate, and visualize the split that each of these methods would choose. (3 pts)

(b) Use the implementation of random forests in sklearn[1] to classify the jet tagging data. Perform the following steps:

i) Load the data and split it into train, validation and test set. Validation and test set should each contain $N = 200$ data points with the rest belonging to the training set.

ii) Train the following combination of parameters on the train set and evaluate the learned model on the validation set.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

- Number of trees in $\{5, 10, 20, 100\}$
- Split criterion either Gini or Entropy.
- Depth of the individual trees in $\{2, 5, 10, \text{pure}\}$[2]

iii) Finally choose your preferred set of hyperparameters and evaluate the performance on the test set.

(4 pts)

4 Beta Distribution

Consider a binary classification problem on one-dimensional data $(x_i, y_i), i = 1 \dots N, x_i \in \mathbb{R}, y_i \in \{0, 1\}$ with labels l_i. For every location $x \in \mathbb{R}$ we model the probability of it belonging to class 1 with a Bernoulli distribution with mean μ_x. To estimate μ_x, we consider data points in the vicinity of x, more precisely we count for each of the two classes all points of that class within distance r of x, $c_{x,y} = |\{i | y_i = y, \, |x - x_i| < r\}|$, and consider those as observations at x.

(a) Under these assumptions, what is the posterior distribution of μ_x? (1 pt)

(b) For the data given in `data1d.npy` with labels given in `labels1d.npy`, $r = 0.3$ and $x = -5, -4.99, \dots, 4.99, 5$, compute the counts $c_{x,y}$. (3 pts)

(c) For each x as in (b), evaluate the probability density function of the posterior distribution of μ_x at $\mu_x = 0, 0.01, \dots, 0.99, 1$, and save the results in a matrix M. Visualize this matrix as an image, and interpret the results. (3 pts)

(d) Repeat the above for smaller and larger values of r and interpret the results. Which r would you choose? (1 pt)

(e) Bonus Instead of counting the points inside a certain distance, use a radial basis function kernel

$$k(x - x_i; \sigma) \coloneqq \exp\left(-\frac{(x - x_i)^2}{2\sigma^2}\right)$$

to assign fractional counts $k(x - x_i)$ to points x_i based on their distance to x. Implement this and repeat parts (b), (c) and (d) in this setting (with the bandwidth σ taking the role of r). (3 pts)

5 Bonus: The Multivariate Normal

Consider a two-dimensional Normal distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \\ &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),\end{aligned}$$

formulated once with the covariance matrix $\boldsymbol{\Sigma}$ and once with the precision matrix $\boldsymbol{\Lambda}$, where $\mathbf{x} = (x_1, x_2)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

Derive that $p(x_1 | x_2 = c) = \mathcal{N}(x_1 | \mu_{1|2}, \Sigma_{1|2})$ and give the expressions for $\mu_{1|2}$ and $\Sigma_{1|2}$. To get from $p(\mathbf{x}) = p(x_1, x_2)$ to the conditional we can just fix x_2 to the observed value c and normalize the expression. In order to do this go through the following steps:

[2] where pure refers to growing each tree until each leaf is pure

1. Consider $p(\mathbf{x})$ and, ignoring the normalization constant, expand the square in the exponential sorting it into terms depending on x_1 and those independent of it. Do this in the form of the $\mathbf{\Lambda}$ instead of $\mathbf{\Sigma}$ for simplicity.

2. The resulting term is again quadratic, i.e. has the form of a Gaussian and you only need to find $\mu_{1|2}$ and $\Sigma_{1|2}$. Do this by comparing the form you get via 1. with the expanded exponent of a general Gaussian, comparing the relevant coefficients in each term. This allows you to write $\mu_{1|2}$ and $\Sigma_{1|2}$ in terms of $x_2, \mu_1, \mu_2, \Lambda_{11}, \Lambda_{12}$.

3. It can be shown that

$$\begin{aligned} \Lambda_{11} &= \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} \\ \Lambda_{12} &= -\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}. \end{aligned}$$

   Use these results to finally formulate $\mu_{1|2}$ and $\Sigma_{1|2}$ in terms of $x_2, \mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{21}$.

(3 pts)