Spring 2024

ESE 588 – Project

Your report should be presented in one PDF file with a maximum of 20 single-spaced pages (11 or 12 pt), including tables and figures. Do not upload any code.

Part 1 - Bayesian linear regression

A. Sequential Bayesian learning

Generate synthetic data according to

$$y_n = \theta_0 + \theta_1 x_n + \eta_n,$$

where $x_n \sim \mathcal{U}(-1, 1)$, $\eta_n \sim \mathcal{N}(0, \sigma_\eta^2)$, $\theta_0 = 1$, and $\theta_1 = -1$, $n = 1, 2, \cdots, 10$.

1. Set $\sigma_\eta^2 = 1$ and assume that it is known. Let $\boldsymbol{\theta}$ have a Gaussian prior $\mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$, where $\sigma_\theta^2 = 4$ and is also assumed known. Produce contour plots of the posterior of $\boldsymbol{\theta}$ after
   (a) x_1 and y_1 are observed,
   (b) after one more pair of points is observed, i.e., after x_2 and y_2 also become available, and
   (c) after 8 more pairs of points are observed, i.e., find the posterior given $\{x_n, y_n\}_{n=1}^{10}$, $p(\boldsymbol{\theta}|x_{1:10}, y_{1:10})$.

   Discuss briefly the obtained results.
2. For each case (a)–(c), draw 10 pairs of $\boldsymbol{\theta}$ from the respective posteriors and plot the obtained lines. Namely, draw 10 values of $\boldsymbol{\theta}$ from the posteriors and for all the drawn values plot the lines defined by these values on a same figure.
3. Use the first three pairs $\{x_n, y_n\}_{n=1}^3$ to find the predictive distribution $p(y_*|x_*, x_1, t_1, x_2, t_2, x_3, t_3)$, where x_*, takes values from -1 to 1 in steps of 0.01, and plot the mean of the predictive distribution with one standard deviation on either side of the mean. Discuss briefly the obtained results.
4. Repeat 1., 2., and 3. with $(\sigma_\eta^2 = 5, \sigma_\theta^2 = 4)$ and $(\sigma_\eta^2 = 1, \sigma_\theta^2 = 25)$. Discuss the difference with respect to the results from 1. and 2. and 3.

B. Gaussian basis functions

Define the Gaussian basis functions

$$\phi_n(x) = \exp(-(x - x_n)^2/2s^2), \tag{1}$$

for $n = 1, 2, \cdots, N$, and where the means x_n are known. Generate data according to

$$y_n = x_n^2 + \eta_n, \tag{2}$$

where $x_n \sim \mathcal{U}(-5, 5)$, $\eta_n \sim \mathcal{N}(0, \sigma_\eta^2)$, where $\sigma_\eta^2 = 1$. For modeling the signal x^2, we use

$$y(x, \theta) = \sum_{n=1}^{N} \theta_n \phi_n(x), \tag{3}$$

where the coefficients θ_n are unknown but have prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$, where $\sigma_\theta^2 = 9$.

1. Set $s = 1$ and use the first three generated pairs $\{x_n, y_n\}_{n=1}^3$ to find the predictive distribution $p(y_*|x_*, x_1, y_1, x_2, y_2, x_3, y_3)$, where x_*, takes values from -1 to 1 in steps of 0.01. Plot the mean of the predictive distribution with one standard deviation on either side of the mean. Discuss briefly the obtained results.

2. Repeat with $s = 0.5$ and $s = 0.1$ and discuss the differences in prediction.

3. Repeat 1. with $\{x_n, y_n\}_{n=1}^{10}$. Compare the results with those from 1.

Part 2 - Classification

Consider a two-class, two-dimensional classification problem, where the first class, ω_1 is modeled by a Gaussian distribution with a mean and covariance given by

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

and where the second class (ω_2) is modeled by a Gaussian distribution with a mean and covariance given by

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 2.5 \\ 2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.25 & -0.3 \\ -0.3 & 1 \end{bmatrix}$$

1. Form and plot a data set $\mathcal{X}$ consisting from 200 points from ω_1 and another 200 points from ω_2.

2. Use the first 100 points to learn the mean and covariance of the respective bivariate Gaussians and then assign each one of the points of the remaining samples from $\mathcal{X}$ to either ω_1 or ω_2 according to the Bayes' decision rule. Plot the points with different colors, according to the class they are assigned to, find the confusion matrix of the classifier, and compute the classification error probability. Explain how the classification is carried out.

3. Classify the generated data with logistic regression.

4. Repeat 1 and 2 with 1000 points generated from each class. Use half of the points for training.

5. Write a brief conclusion from all the results.

Part 3 - Classification of real data

The UCI Machine Learning Repository hosts a dataset comprising measurements of fetal heart rate (FHR) and uterine contraction (UC) features extracted from cardiotocograms (CTGs). Expert obstetricians classify the CTGs into one of three categories reflecting the fetus's well-being: Normal (coded as 1), Suspect (coded as 2), and Pathologic (coded as 3). The dataset consists of 21 features derived from these measurements. You can access the dataset at the following link:

https://archive.ics.uci.edu/dataset/193/cardiotocography

In the spreadsheet, the features are listed from columns K to column AE.

1. Use two machine learning methods of your choice for classification and compare their performance. Ensure that the comparison is fair by using the same datasets for both training and testing.

2. In a paragraph, explain the selected methods (with references) and justify your choice.

3. Discuss the results comparing the performance of the methods along with specifics of their training and testing.