Indian Institute of Technology, Jodhpur

Machine Learning - 1 | Assignment-2

Due Date: Theory Questions: 17/3/3021, Programming Questions: 28/03/2021,

---

Submission Policy :

- Any kind of plagiarism is not accepted. We will strictly follow the institute policies for plagiarism.
- Recommended Programming language : Python + Pandas (for reading csv file) + Sklearn
- Submission should include: working code for all parts and a single pdf report for all questions,
- Submit a single zip file that contains:
    1. Code for question 1 and 2
    2. A single pdf report for all questions. Report must contain all the results obtained from codes
- The theory assignment should include proper explanation and full solution of theoretical questions.

Question 1 **30 Marks**

Dataset : Iris Dataset

Use sklearn library for loading iris dataset.

Aim: Classification using Naive Bayes classifier

- Apply Naive bayes classifier assuming all features are independent.

Do not use any predefined library for classification

Report overall accuracy, class wise accuracy, confusion matrix and ROC curve.

Question 2 **60 marks**

Dataset: Wine dataset

(use sklearn library for loading the dataset)

Aim: Naive Bayes Classification

Shuffle the data with seed value 42 and perform a 70- 30 stratified split of the data into a train and test set. Also, plot the class-wise distribution of data in the train and test set (one for train set and one for test set). Compare the distributions. Now, perform classification as follows: **10 + (2*5) = 20 marks**

(a) Train a Gaussian Naive Bayes classifier and report (a) the class priors, (b) mean and variance of each feature per class.

(b) Train another Gaussian Naive Bayes classifier by setting prior probability for the classes. Repeat this experiment by setting priors in the ratios: (a) 40-40-20 and (b) 80-10-10.

For all the experiments above, report the (a) accuracy, (b) confusion matrix, on the train and test set. State your observations and analysis. The report should include the distributions as well.

10 + (10*3) = 40 marks

Theory:

Question 3 **40 Marks**

Write your answers for the following question in the report.

(a) What is Naive Bayes' assumption? How does it help? Explain with an example.

(b) For a Gaussian Bayes classifier, which one of these structural assumptions is the one that most affects the trade-off between underfitting and overfitting:

(i) Whether we learn the class centers by Maximum Likelihood or Gradient Descent

(ii) Whether we assume full class covariance matrices or diagonal class covariance matrices.

(iii) Whether we have equal class priors or priors estimated from the data.

(iv) Whether we allow classes to have different mean vectors or we force them to share the same mean vector

(c) Consider a two-class one-feature classification problem with the following Gaussian class-conditional densities. $p(x|w1) = N(0, 1)$ and $p(x|w2) = N(1, 2)$.

Assume equal prior probabilities and 0-1 loss function. Solve the following two questions and show your work.

(i) What is the Bayes decision boundary?

(ii) Suppose the prior probabilities are changed as follows: $P(w1) = 0.6$ and $P(w2) = 0.4$. How will the decision boundary change?

(d) Consider the multivariate normal density for which $\sigma_{ij} = 0$ and $\sigma_{ii} = \sigma_i^2$, i.e., $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$.

(i) Show that the evidence is

$$p(\mathbf{x}) = \frac{1}{\prod_{i=1}^{d} \sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right].$$

(ii) Plot and describe the contours of constant density.

---------- **** ----------