Assignment 1: Bayesian Decision Theory and Parametric Estimation

Submission: Friday April 22th
Groups of maximum 3 students

Prof. Fabio A. González
Machine Learning - 2022-I
Maestría en Ing. de Sistemas y Computación

---

1. (1.0)

   (a) Download the Pima Indians Diabetes dataset. Load the dataset as a numpy array. The `outcome` column corresponds to the class label, the other columns to input features.

   (b) Split the dataset in training and validation using the following `scikit-learn` command:

   ```
   train_test_split(X, y, test_size=0.33, random_state=42)
   ```

   where `X` and `y` are the input features and labels respectively.

   (c) For each one of the 8 input features build a univariate Gaussian classifier estimating the parameters using the training dataset and evaluating the classifier in the validation dataset. For each classifier write down the parameters. Report accuracy, precision and recall. Which classifier is the best? For this classifier draw the curves for the posterior distribution for each class and show where the decision boundary is.

2. (1.0)

   (a) In this question, you will take all the combinations of pair of features. For each pair of features you will model the classes as bivariate Gaussian distributions with a covariance matrix $\Sigma = I\sigma$, where σ is a scalar, shared by all the classes. This is, the distribution for each class has a different mean but the same covariance matrix.

   (b) For each combination estimate the parameters using the training dataset and evaluate the classifier in the validation dataset. Which combination of parameters obtained the best performance? Report the evaluation results for the best performing combination.

   (c) Draw ROC curves for the best performing univariate model (from question 1) and bivariate model. Which model is better? Explain the results.

3. (1.0)

   (a) For the best bivariate model (from question 2) draw a plot, where the regions corresponding to the two classes are shown with different colors. A region corresponding to a class is the set of points where the particular class discriminant function is maximum (decision regions, [Alp14] Sect. 3.4).

   (b) The boundary between class regions must be a line. Calculate the equation of this line clearly explaining the deduction process. Draw the line along with the regions.

4. (1.0)

   (a) Repeat the previous question, but this time build a model that minimizes the risk. In this case consider that the cost of a false positive is 2, the cost of a false negative is 1 and the cost of a correct classification is 0. Compare with the previous model. Discuss.

   (b) Calculate the ROC curve for this model and compare with the ROC curves from question 2. Discuss.

5. (1.0)

   (a) Using the best combination of parameters found in question 2 build a bivariate classifier, but this time the covariance matrix could be an arbitrary matrix (not diagonal) and different for each class.

   (b) Draw the regions corresponding to each class. Compare with the regions obtained in question 3. Discuss.

   (c) Add the possibility of rejection to your model. Draw the three regions corresponding to both classes and rejection. Discuss.

The assignment must be submitted as a Jupyter notebook through the following Dropbox file request, before midnight of the deadline date. Assignments submitted after this time will be ignored. Before submitting be sure that the notebook runs and that all figures are correctly rendered. The file must be named as `ml-assign1-unalusername1-unalusername2-unalusername3.ipynb`, where `unalusername` is the user name assigned by the university (include the usernames of all the members of the group). Do not submit additional files, only the notebook file.

References

[Alp14] Alpaydin, E. Introduction to Machine Learning, 3Ed. The MIT Press, 2014

[DHS00] Duda, R. O., Hart, P. E., and Stork, D. G. 2000 Pattern Classification (2nd Edition). Wiley-Interscience.