Exercise Set 5

Please write the name(s) of the student(s) to the answer sheet! Submission DL 18 May 9:15. The solutions will be presented in a session at noon 18 May, where you should be present (you may be asked to present your solutions).

Problem 1

Learning objective: Naive Bayes classifier, binary probability distributions, probability computations in classification.

Use the discrete naive Bayes classifier in the following problem. This problem should be solvable by pen, paper, and your pocket calculator. Assume that the class variable is binary Y, and that there are three input features X_1, X_2, and X_3 which are also binary.

The training data is as follows:

| Y	X_1	X_2	X_3
0	1	1	0
0	0	1	1
0	0	0	0
1	1	0	0
0	1	0	0

Tasks

- Provide estimates of the class-conditional distributions of the form $P(X_j = x \mid Y = c)$ for all j, x, and c without smoothing. *Hint:* Use empirical frequencies.
- Do the same with Laplace smoothing (https://en.wikipedia.org/wiki/Additive_smoothing) by using pseudocounts $m_{c,j,x} = 1$ added to the empirical frequencies. (Recall: You can interpret the pseudocounts as effects of the prior, see E2P1, i.e., Laplace smoothed frequencies can be though as MAP estimates!)
- Assume that the required probabilities have been estimated from data. How is the class value predicted for a new test instance (x_1, x_2, x_3)? Provide a formula. How would you classify a test instance $(1, 0, 0)$? (Remember to apply smoothing to the class probabilities $P(Y = c)$, too!)

Problem 2

Learning objectives: basic principles of decision trees.

Familiarise yourself with Section 8.1 of James et al. Pick one of the impurity measures presented in Equations (8.5), (8.6), or (8.7). Then simulate the tree building algorithm by hand.
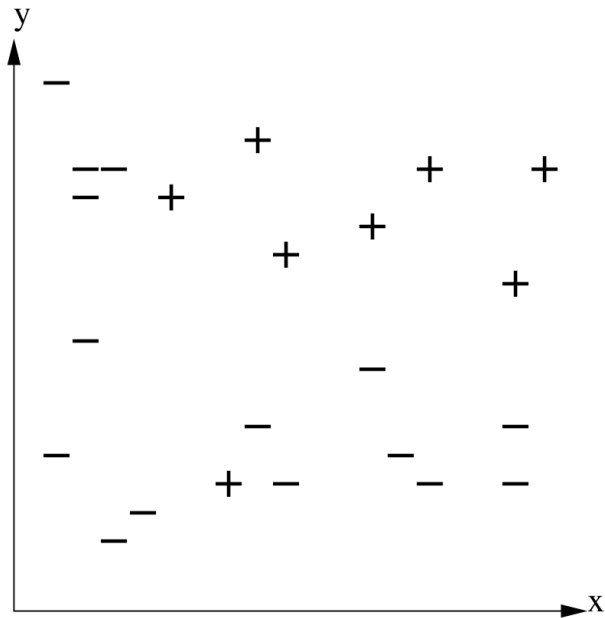
Tasks

- Sketch a run of the classification tree algorithm with the toy data set in the Figure above (binary classification task in $\mathbb{R}^2$) and draw the resulting classification tree. Report the values of the chosen impurity measure for each split and try to choose the splits that obtain the best impurity measure. You do not need to worry about overfitting here: the resulting classification tree can fit the training data with no error. Don't worry that you don't count the classes exactly right or that your results are not super-accurate as long as they are "in the ballpark".

Problem 3

Learning objectives: basics of k-NN method.

Consider the problem of applying *k-nearest neighbour* (k-NN) classifier to the training dataset $D = \{(x_i, c_i)\}_{i=1}^{14}$, where the covariates $x_i \in \mathbb{R}$ and the classes $c_i \in \{-1, +1\}$ are given in below. You should be able to do this by pen and paper.

i	x_i	c_i
1	0	+1
2	2	+1
3	3	+1
4	5	−1
5	6	+1
6	8	+1
7	9	+1
8	12	−1
9	13	−1
10	15	−1
11	16	+1
12	18	−1
13	19	−1
14	21	−1

Tasks

- Where are the classification boundaries for the 1-NN and 3-NN classifiers? What are the respective classification errors on the training dataset?
- How does the choice of k in k-NN affect the classification boundary (not in the above example but in general)? Give examples of the behaviour for extreme choices.

Problem 4 (bonus)

Learning objectives: Understanding discriminative vs. generative learning.

[This problem is optional. Do this if you have time.]

Download Ng et al. (2001). You **do not need to read the full paper** in detail or understand all of the details! Rather try to find the answers to the following questions.

Reference: Ng, Jordan (2001) On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. NIPS. http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf (http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf)

Tasks

- Read the *abstract* and the *Introduction*. Is discriminative learning better than generative learning, according to the authors? Justify your answer.
- By a "parametric family of probabilistic models", the authors mean a set of distributions, where each distribution is defined by a set of parameters. An example of such a family is our friend, the family of normal distributions where the parameters are μ and Σ. Ng and Jordan denote by h_{Gen} and h_{Dis} two models chosen by optimizing different things. Find an explanation of what these "things" are that are being optimized in the paper, i.e., what characterizes these two models. Which two families do the authors discuss, and what are the (h_{Gen}, h_{Dis}) pairs for those models?

- Study Figure 1 in the paper. Explain what it suggests (see the last paragraph of the Introduction). Reflect this on the previous item.

Optional reading

You will not get any points for reading these articles, but to conclude discussion of the classifiers you can, if you want, take a look at the following review:

- Fernández-Delgado et al. (2014) Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?. JMLR. http://jmlr.org/papers/v15/delgado14a.html (http://jmlr.org/papers/v15/delgado14a.html)

Hand's article is still timely:

- Hand (2006) Classifier Technology and the Illusion of Progress. Statistical Science. https://doi.org/10.1214/088342306000000060 (https://doi.org/10.1214/088342306000000060)

In [ ]: