Elements of Machine Learning, WS 2022/2023
Jilles Vreeken and Aleksandar Bojchevski
Exercise Sheet #2: *Classification*

Deadline: Thursday, December 1, 2022, 16:00

Before solving the exercises, read the instructions on the course website.

- For each theoretical problem, submit a single `pdf` file that contains your answer to the respective problem. This file may be a scan of your (legible) handwriting.
- For each practical problem, submit a single `zip` file that contains
  - the completed jupyter notebook (`.ipynb`) file,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows all your results.
- For the bonus question, submit a single `zip file` that contains
  - a `pdf` file that includes your answers to the theoretical part,
  - the completed jupyter notebook (`.ipynb`) file for the practical component,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows your results.
- Every team member has to submit a signed Code of Conduct.

Problem 1 (T, 8 Points). **Logistic regression.**

1. [4*pts*] In which setting is logistic regression applicable? Explain at least three problems with linear regression when applied in such a setting.
2. [1*pts*] What do we model with logistic regression? How are the independent variables and obtained probabilities related?
3. [1*pts*] In general, what is the meaning of odds? Write down the formula and explain in your own words. How do odds relate to logistic regression?
4. [1*pts*] Let X be a scalar random variable. Prove that

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \iff \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \; .$$

   What is the relationship between the logistic and the logit function? Why is this information about the relationship important? Explain.

5. [1*pts*] Let Y_θ be a binary random variable for which

$$\mathbb{P}(Y_\theta = 1) = \frac{e^\theta}{1+e^\theta} \;, \qquad \text{where } \theta \in \mathbb{R} \;,$$

   and define a parameter vector $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]$ and feature vector $\boldsymbol{x} = [1, x_1, \ldots, x_p]$. Show that

$$\frac{odds(Y_{\boldsymbol{x}^\top \boldsymbol{\beta} + \beta_i \delta})}{odds(Y_{\boldsymbol{x}^\top \boldsymbol{\beta}})} = \exp(\beta_i \delta) \;, \qquad \text{for some } \delta \in \mathbb{R} \text{ and any } i \in \{1, \ldots, p\}$$

   and explain the meaning of this equality in your own words.

Problem 2 (T, 10 Points). **Bayes-optimal classifier.** The optimal misclassification error is achieved by the Bayes optimal classifier. This is the classifier that assigns every point X to its most likely class. That is, the Bayes optimal classifier predicts

$$\hat{y} = f^*(x) = \arg\max_{y \in \{0,1\}} P(Y = y | X = x) \; .$$

1. Consider a scalar feature $X \in \mathbb{R}^2$ and a binary random variable Y, for which

$$P(X|Y=0) = \begin{cases} \frac{1}{\pi r^2} & \|X\| \leq r \\ 0 & \text{otherwise} \end{cases} \; , \qquad \text{and}$$

$$P(X|Y=1) = \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2 \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\|X\|^2}{2\sigma^2}\right) \; , \qquad \text{with}$$

$$P(Y=0) = cP(Y=1) \; ,$$

where $r, \sigma > 0$ and $0 < c < 1$ are parameters.

   (a) [6*pts*] Derive the Bayes optimal classifier for Y as a function of r and σ.

   (b) [2*pts*] Draw the decision boundary for $\sigma = 1$, $r = e\sqrt{2} \approx 3.84$ and $c = \exp(-\frac{1}{3})$; explain your observations. What will happen to the decision boundary, if we increase c while keeping all the other parameters fixed?

   Note: For this, you have to find the region of $\mathbb{R}^2$ for which $P(Y = 1|X) \geq P(Y = 0|X)$.
   Hint: Use the Bayes formula given in the lecture.

2. [2*pts*] Given that the Bayes optimal classifier has the lowest misclassification error among all classifiers, why do we need any other classification method?

Problem 3 (T, 5 Points). **So Many Classifiers.** We now know four different classifiers: K-NN, LDA, QDA, and Logistic Regression (LR).

1. [4*pts*] Which assumptions do each of the models make w.r.t. the data distribution? Depending on the type of decision boundary, which of the respective methods would you recommend?

2. ([1*pts*] Although LDA and LR often yield similar results, LR is often preferred. Give two reasons for this.

Problem 4 (P, 19 Points). **Speech Recognition.** We will now consider LDA and QDA for a real-world speech recognition task. The data we consider contains digitized pronunciation of five phonemes: `sh` as in “she”, `dcl` as in “dark”, `iy` as the vowel in “she”, `aa` as the vowel in “dark”, and `ao` as the first vowel in “water”. These phonemes correspond to responses/classes (column name `g`). The dataset contains 256 predictors (log-periodograms, which is a common way of representing voice recordings in speech recognition).
Use `Practical_Problem_1.ipynb` found in the `a1_programming` file from the course website.

1. [1*pts*] Load the phoneme data set `phoneme.csv` and split the dataset into a training and test set according to the `speaker` column. Then exclude the `row.names`, `speaker` and response column `g` from the features.

2. [2*pts*] Fit an LDA model to classify the response based on the predictors; then compute and report train and test error. *Useful functions:* `sklearn.model_selection.StratifiedShuffleSplit`.

3. [3*pts*] Plot the projection of the training data onto the first two canonical coordinates of the LDA.

4. [$4pts$] Select the two phonemes `aa` and `ao`. Fit an LDA model on this data set and repeat the steps done in (2).
5. [$6pts$] Repeat steps (2) and (4) using QDA and report your findings. Would you prefer LDA or QDA in this example? Why?
6. [$3pts$] Generate confusion matrices for the LDA and QDA model for `aa` and `ao`. Which differences can you observe between the models?

Problem 5 (Bonus). **Shattering Data.**
This bonus problem contains both theoretical and practical parts.

1. **Theory**. First, we dive into the classification flexibility of classifiers.

   We first define the ability of a family $\mathcal{F}$ of classifiers to "shatter" a set of points $\boldsymbol{X}$, that is to correctly classify them into two classes, for any possible assignment of binary labels to them. The maximum number of distinct points that can be shattered by at least one member of a classifier in the family is called the Vapnik Chervonenkis (VC) dimension of the family.
   Formally, for a family of classifiers $\mathcal{F}$ that can classify points that lie on some domain D we define

   $$VC(\mathcal{F}; D) = \max\left\{k \in \mathbb{N} \,\middle|\, (\exists X \subset D), |X| = k : (\forall S \subseteq X)\,(\exists f \in \mathcal{F}) \text{ for which } S = \{x \in X \mid f(x) \geq 0\}\right\}.$$

   (a) [$1pts$] Show that $VC(\mathcal{F}_{\text{LC}}; \mathbb{R}^2) = 3$, where $\mathcal{F}_{\text{LC}}$ is the family of all linear classifiers over two features, without allowing any interactions.
   For this, you have to find any example of 3 points which can be shattered, but also prove that no set of 4 points can be shattered.

   (b) [$1pts$] Show that $VC(\mathcal{F}_{\text{LC}}; \mathbb{R}^3) \geq 4$.

   (c) [$1pts$] Show that $VC(\mathcal{F}_{\text{QDA}}; \mathbb{R}^2) \geq 6$, where $\mathcal{F}_{\text{QDA}}$ is the family of all QDA classifiers.

2. **Practical**. Consider the notebook `Bonus_Problem.ipynb`.

   (a) Open the dataset `data_1` and study its distribution. Based on the intuition you acquired in the theoretical part of this problem, can it be classified sufficiently well with a linear classifier?

   (b) You will now modify the feature vectors of the observations in this dataset so that it can be classified with a linear classifier. To do so

   - create at most two additional dummy variables (features) based on the original ones
   - apply logistic regression on the derived feature vectors
   - plot the distribution of the test dataset alongside the decision boundary of your classifier
   - compute and measure your mis-classification error.

   Explain your observations.

   (c) Open the dataset `data_2` and study its distribution. Based on the intuition you acquired in the theoretical part of this problem, can it be classified sufficiently well with a linear classifier?

   (d) You will now modify the feature vectors of the observations in this dataset so that it can be classified with a linear classifier. To do so

   - transform the original feature vectors to form a single-dimensional feature,
   - apply logistic regression on the derived feature vectors
   - plot the distribution of the test dataset alongside the decision boundary of your classifier
   - compute and measure your mis-classification error.

   Explain your observations.