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

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

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

Solution.

1. If the response variable is **categorical/qualitative**, logistic regression is applicable. Contrast this to linear regression, which, in general, is applicable if the response variable is **quantitative**. (1 Point) *Even though, we can transform the categorical response variable via dummy variable to a quantitative variable, the corresponding linear regression model will suffer from the following problems:*

   (a) **implicit unnatural ordering:** a number of classes $K \geq 3$ introduces an undesired ordering and distances on the classes that will be respected by the regression the model depends largely on this ordering of, and the distances between, the numbers representing the categories (1 Point)

   (b) **masking effect:** when the number of classes K exceeds two, especially when K is large, the rigid nature of the linear regression model may lead to one or more classes being masked (always dominated, hidden) by others (1 Point)

   (c) **out of probability range:** estimators can be negative or greater than 1, as a consequence of the rigid nature of linear regression. This breaks the interpretability of the response as a probability. (1 Point)

   Additional/optional answer: response variable is not normally distributed.
   Grading: 0.5 Points for mentioning, 0.5 Points for explanation. Italic text not required for full points.

2. In logistic regression the output of a model is the probability $\hat{p}(\boldsymbol{x})$ that Y belongs to a particular class k, having seen a given feature vector $\boldsymbol{x} = (1, x_1, \ldots, x_p)$,

$$p(\boldsymbol{x}) = \mathbb{P}(Y = k | X = \boldsymbol{x})\,, \quad [0.5pts]$$

   under the assumption that the logistic model is true.

   According to the logistic model, this probability equals the logistic function f of a linear combination $\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}$ of the extended features $\tilde{\boldsymbol{x}} = (1, x_1, \ldots, x_p)$,

$$p(\boldsymbol{x}) = f\left(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}\right)\,, \qquad [0.5pts]$$

   where f is the logistic function

$$f(z) = \frac{e^z}{1 + e^z}\,. \tag{1.1}$$

   Thus, overall, the complete probability modelled by the logistic regression becomes

$$p_{\boldsymbol{\beta}}(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}}}{1 + e^{\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}}}$$

   and is parameterised by the linear coefficients $\boldsymbol{\beta}$.

3. The odds of a binary random variable is the probability of its value $X = 1$ over that of its other value $X = 0$; that is,

$$odds(x) = \frac{p(X = 1)}{p(X = 0)}\,. \qquad [0.5pts]$$

   Logistic regression is essentially a linear model whose response is the log-odds (i.e., the natural logarithm of the odds) of the class variable Y,

$$\ln\left(\frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)}\right) = \boldsymbol{X}\boldsymbol{\beta}\,.$$

   The importance of this relation lies in the following observation. Since the logistic function of Eq. (1.1) is the inverse of the log-odds function, we can use the logistic of the linear combination $\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}$ to compute the class probability $\mathbb{P}(Y = 1)$. $[0.5pts]$

4.

$$
\begin{aligned}
p(X) &= \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} && \mid *(1+e^{\beta_0+\beta_1 X})\\
\Leftrightarrow p(X)(1+e^{\beta_0+\beta_1 x}) &= e^{\beta_0+\beta_1 x} && \mid \text{distributive law}\\
\Leftrightarrow p(X)+p(X)e^{\beta_0+\beta_1 x} &= e^{\beta_0+\beta_1 x} && \mid -p(X)e^{\beta_0+\beta_1 x}\\
\Leftrightarrow p(X) &= e^{\beta_0+\beta_1 x} - p(X)e^{\beta_0+\beta_1 x} && \mid \text{ditributive law}\\
\Leftrightarrow p(X) &= e^{\beta_0+\beta_1 x}(1-p(X)) && \mid : (1-p(X))\\
\Leftrightarrow \frac{p(X)}{1-p(X)} &= e^{\beta_0+\beta_1 x} && \mid ln\\
\Leftrightarrow ln(\frac{p(X)}{1-p(X)}) &= \beta_0+\beta_1 x \quad \text{(1 Point)}
\end{aligned}
$$

The logit function is the inverse function of the logistic function (0.5 Point):

$$
\begin{aligned}
&f : x' \longrightarrow y\\
&g : y \longrightarrow x' \quad \text{(inverse function)}\\
&y = f(x') = \frac{e(x')}{1-e(x')}\\
&x' = g(y) = log\left(\frac{y}{1-y}\right)\\
&x' = g(f(x')) \quad \text{(inverse)}\\
&\text{here: } x' = \beta_0+\beta_1 x \in (-\infty,+\infty), \quad y = p(x) \in (0,1)
\end{aligned}
$$

The logit function maps probabilities from the range (0,1) to the entire real number range $(-\infty,\infty)$ *(where it is usually the logarithm of the odds)*, acts as a link function for logistic regression and provides the relationship between the linear predictor and the outputted probability. $[0.5pts]$

5. To show the requested relation we first observe that the odds of the random variable are equal to e^{θ}:

$$
odds(Y_\theta) = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} = \frac{\frac{e^\theta}{1+e^\theta}}{1-\frac{e^\theta}{1+e^\theta}} = \frac{\frac{e^\theta}{\cancel{1+e^\theta}}}{\frac{1+\cancel{e^\theta}-\cancel{e^\theta}}{\cancel{1+e^\theta}}} = e^\theta\,.
$$

Therefore, for the equation we need to show, we have

$$
\frac{odds(Y_{\boldsymbol{x}^\top\boldsymbol{\beta}+\beta_i\delta})}{odds(Y_{\boldsymbol{x}^\top\boldsymbol{\beta}})} = \frac{e^{\boldsymbol{x}^\top\boldsymbol{\beta}+\beta_i\delta}}{e^{\boldsymbol{x}^\top\boldsymbol{\beta}}} = e^{\beta_i\delta}\,. \qquad [0.5pts] \tag{1.2}
$$

To interpret this result, we first observe that adding $\beta_i\delta$ to $\boldsymbol{x}^\top\boldsymbol{\beta}$ is equivalent to changing the value of predictor variable x_i from x_i to $x_i+\delta$. We can now see that, when the value of a predictor changes by δ, this modifies the new odds-ratio by a multiplicative factor of $\exp(\beta_i\delta)$. $[0.5pts]$

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

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

Solution.

1. (a) For this we can express the density function of the uniform distribution using the indicator function as

$$\mathbb{1}_{\{\|X\|\leq r\}} := \begin{cases} 1 & \|X\| \leq r \\ 0 & \text{otherwise} \end{cases}\ ,$$

so that the probability density of X under $Y = 0$ becomes

$$p(X|Y=0) = \frac{1}{\pi r^2}\mathbb{1}_{\{\|X\|\leq r\}}\ .$$

We need to solve

$$P(Y=0|X) \geq P(Y=1|X) \iff \frac{P(X|Y=0)}{\cancel{P(X)}P(Y=0)} \geq \frac{P(X|Y=1)}{\cancel{P(X)}P(Y=1)} \iff$$

$$\frac{P(X|Y=0)}{P(X|Y=1)} \geq \frac{P(Y=0)}{P(Y=1)} \iff$$

$$\frac{2\cancel{\pi}\sigma^2}{\cancel{\pi}r^2}\,\frac{\mathbb{1}_{\{\|X\|\leq r\}}}{\exp(-\frac{1}{2\sigma^2}\|X\|^2)} \geq c \overset{\log(\cdot)}{\iff}$$

$$\log\left(\frac{2\sigma^2}{r^2}\right) + \log\left(\mathbb{1}_{\{\|X\|\leq r\}}\right) + \frac{1}{2\sigma^2}\|X\|^2 \geq \log c \iff$$

$$\|X\|^2 \geq 2\sigma^2\left[\log c - \log\left(\frac{2\sigma^2}{r^2}\right) - \log\left(\mathbb{1}_{\{\|X\|\leq r\}}\right)\right]$$

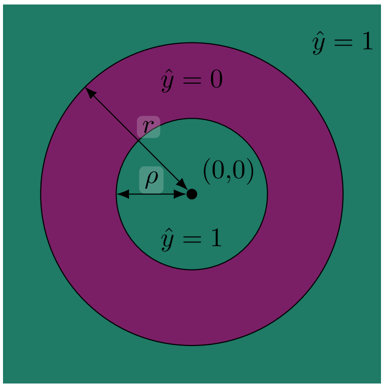
We now take 2 cases.

- When $\|X\| > r$, we have $\mathbb{1}_{\{\|X\| \leq r\}} = 0$, so that its negative logarithm explodes to infinity, and the original inequality is impossible. Then, the Bayes optimal classifier always predicts $\hat{y} = 1$.
- When $\|X\| \leq r$ the indicator function becomes 1, so that its logarithm vanishes. Then we get:

$$P(Y = 0|X) \geq P(Y = 1|X) \iff \|X\|^2 \geq 2\sigma^2 \left[\log\left(\frac{2\sigma^2}{r^2}\right) + \log c\right] \overset{\sqrt{\cdot}}{\iff}$$

$$\|X\| \geq \sqrt{\max\left(0, 2\sigma^2 \left[\log c - \log\left(\frac{2\sigma^2}{r^2}\right)\right]\right)} =: \rho$$

which describes a circle of radius ρ.

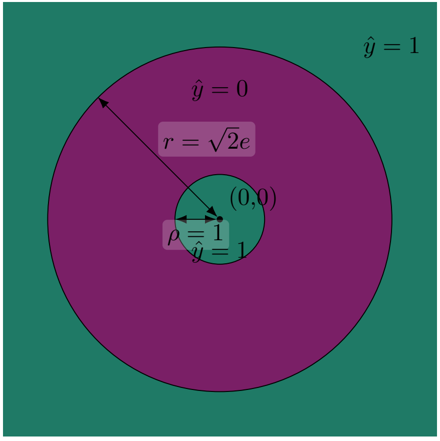


Overall, whenever $\rho < r$ we get the decision boundary which contains a ring-like region for which the classifier predicts $\hat{y} = 1$, and for all the other parts decides for $\hat{y} = 1$.
Whenever $\rho > r$, the classifier always predicts $\hat{y} = 1$.

(b) We now replace the parameters we are given in the decision boundary that we found above. For the quantity ρ we compute

$$\begin{aligned}
\rho &:= \sqrt{\max\left(0, 2\sigma^2 \left[\log c - \log\left(\frac{2\sigma^2}{r^2}\right)\right]\right)} \\
&= \sqrt{\max\left(0, 2 \cdot 1^2 \left[\log \exp\left(-\frac{3}{2}\right) - \log\left(\frac{2 \cdot 1^2}{\left(\sqrt{2}e\right)^2}\right)\right]\right)} \\
&= \sqrt{\max\left(0, 2\left[-\frac{3}{2} - \log\left(\frac{2}{2e^2}\right)\right]\right)} = 1
\end{aligned}$$

Therefore, the decision boundary for this case is as shown below.



If the c parameter increases, then the inner radius ρ increases, as we can infer from the formula we computed. The intuition of this is that, since the probability of class $Y = 1$ is higher, we err in favour of this label. Since the probability of the uniform distribution cannot increase outside the outer circle, this only affects the innermost decision boundary.
When this probability increases further, and exceeds the value of $c > \exp\left(-\frac{1}{2}\right)$, the radius of the innermost circle increases and exceeds r. Intuitively, this happens because the prior probability of class $Y = 1$ now dominates that of $Y = 0$, and we therefore always decide for $Y = 1$.

2. Even though the Bayes optimal classifier gives indeed the optimal decision boundary, we often have no access to the actual conditional or prior distributions. These need to be estimated, which introduces an important hurdle that is better to incorporate in the learning method itself.

Even, however, if we could guess or estimate the true parameters of the relevant distributions, often times the computation itself of the decision boundary can be an intractable task. In these cases it sometimes helps to adopt assumptions that allows us to simplify the involved distributions, for instance when we assume the probability of each feature to be independent, given the class, in which case we derive naive Bayes.

As another example, when the true conditional distributions are Gaussian, and for $c = \frac{1}{2}$, the LDA classifier exactly tries to approximate the Bayes optimal classifier. The two coincide in the limit, as the LDA estimates of the conditional distributions converge to the true distribution parameters.

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

1. [$4pts$] 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. ([$1pts$] Although LDA and LR often yield similar results, LR is often preferred. Give two reasons for this.

Solution.

1. - LDA assumes that observations from different classes come from Gaussian distributions with different means and constant variance across classes. It performs well, when the decision boundary is linear. (1 Point)
   - QDA assumes that observations from different classes are come from Gaussian distributions with different means and makes no assumptions about the co-variances. It performs well when the decision boundary is quadratic. (1 Point)
   - LR makes no assumptions on the particular type of data distribution distribution. It performs well, when the decision boundary is linear. *LR produces a probabilistic output.* (1 Point)
   - KNN makes no assumptions about the data distribution within a group (*but assumptions required to define the distance function*). It can fit to an arbitrary shape of decision boundary. *It is a non-parametric method with the difficulty of choosing the correct k. It is sensitive to outliers.* (1 Point)

2. Linear regression requires less assumptions, is more robust (less sensitive to outliers). (0.5 for each of two reasonable reasons)

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. [$1pts$] 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.
   Useful functions: `sklearn.model_selection.StratifiedShuffleSplit`.

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.discriminant_analysis.LinearDiscriminantAnalysis`.

3. [3*pts*] Plot the projection of the training data onto the first two canonical coordinates of the LDA. Investigate the data projected on further dimensions using the `dimen` parameter.

4. [4*pts*] Select the two phonemes `aa` and `ao`. Fit an LDA model on this data set and repeat the steps done in (2).

5. [6*pts*] Repeat steps (2) and (4) using QDA and report your findings. Would you prefer LDA or QDA in this example? Why?

6. [3*pts*] 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) [1*pts*] 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) [1*pts*] Show that $VC(\mathcal{F}_{\text{LC}}; \mathbb{R}^3) \geq 4$.

   (c) [1*pts*] 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.