Proekspert
Spring 2017
Data Science Quiz

Proekspert's Data Scientist test exercise consists of two parts: a quiz (40 % of the points) and an assignment (60 %). The latter is a programming assignment included in this folder, with instructions in DataScienceAssignment.pdf. The former is a self-contained theoretical exercise found in this pdf. Beware: the homework is very difficult. Skilled professionals often cannot expect to solve it for maximum points. Do not be dismayed by this: we will take into consideration your seniority, approach of solution and specific background. Take your time, the whole homework should take about 8 hours.

This quiz contains 13 questions and awards a total of 40 points.

During the quiz, you will tackle a case study of solving an end-to-end machine learning problem. You will demonstrate knowledge of the main concepts of data science. You will reason on a high theoretical level – contrary to the Data Science Assignment, no code is required.

Grade Table

Question	Points	Score
1	2	
2	2	
3	3	
4	2	
5	5	
6	3	
7	6	
8	7	
9	2	
10	4	
11	1	
12	3	
13	0	
Total:	40	

You are approached by a financial lending institution who is interested in predicting which of their clients are going to default on their loans. They present to you a data set consisting of

n=10 000 individuals, for each of them 4 features/measurements being available. These are personal characteristics: age, gender, income, number of children (assume for our purposes that these stay constant). Assume that the financial institution has done some work for you: all running financial measurements (debt outstanding, payments made, payment delays etc) are transformed into one variable: the credit quality rating. This is reassessed every week – per year, we get 52 measurements for each individual. Data was recorded during 2014. The credit quality ratings are then stored in a 10000×52 matrix. Finally, there is a feature called *default* that takes value 1 if a client has defaulted in the next year after the measurements and 0 otherwise.

1. (2 points) Describe the preprocessing steps you would undertake to put the data into a form suitable for modelling.

2. (2 points) Describe in great detail the steps you would undertake in data partitioning to ensure low generalization error of your model.

3. (3 points) Suppose that the number of features is not 4 but 4000. Explain intuitively the concept of the Curse of Dimensionality (CuD). One possible way to alleviate CuD is Principal Component Analysis (PCA). Explain intuitively (but in detail) the process of PCA.

4. (2 points) Suppose that the bank's data engineers screwed up and 80% of the labels have gone missing. This means that for 2 000 individuals, there exists a default measurement (either 0 or 1), but for 8 000 it is missing. How do you proceed?

5. (5 points) One possible way to model the default is with linear regression, i.e. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{y}$ is a n-dimensional vector, $\boldsymbol{X}$ is a $n \times k$ matrix, $\dim(\boldsymbol{\epsilon}) = n$ and n=10 000, k=4.
   (a) (3 points) Give at least two reasons why linear regression is a poor choice if $\boldsymbol{y}$ lies in {0,1}.
   (b) (2 points) Suppose you were predicting a continuous credit quality score instead and using linear regression. Explain why *mean squared error*, defined as $\mathbb{E}[(y-\hat{y})^2]$, is a poor loss/cost function if the real costs are asymmetric: i.e. predicting a too high credit quality score (which ends in default) costs more than predicting a too low credit quality score (which ends in no loans given out).

6. (3 points) It can be proved that choosing the MSE cost function $\mathbb{E}[(y-\hat{y})^2]$ makes the model approximate the conditional expectation, i.e. $\hat{y} \approx \mathbb{E}[y|X]$ while the MAD cost function $\mathbb{E}[|y-\hat{y}|]$ approximates the conditional median $q_{0.5}[y|X]$.
   (a) (1 point) What is the difference between the conditional expectation $\mathbb{E}[y|X]$ and the unconditional expectation $\mathbb{E}[y]$ in our case where y is the default variable.
   (b) (2 points) When can using the MAD cost function be preferred over MSE?

7. (6 points) Suppose you choose logistic regression, defined as $\boldsymbol{y} = \frac{1}{1+e^{-(\boldsymbol{X}\boldsymbol{\beta})}} + \boldsymbol{\epsilon}$
   (a) (2 points) Explain on a high level the maximum likelihood principle to estimating the parameters of the model.

(b) (2 points) Can the parameters $\boldsymbol{\beta}$ be expressed in an analytical form? Why?

(c) (2 points) How would you interpret the feature importances from the trained parameters?

(d) (3 points) (Bonus) You would like to test that the age characteristic has no effect on the default behaviour, i.e. $H_0 : \beta_{age} = 0$. Describe a valid statistical method for testing this in the context of logistic regression.

8. (7 points) Suppose you choose a multilayer perceptron (neural network) as your model, defined as $F(\boldsymbol{X}) = \sigma(\boldsymbol{X}\boldsymbol{\beta})$ where σ is an arbitrary almost surely differentiable non-linear function that may or may not be of the same dimension as $\boldsymbol{X}$

   (a) (3 points) Describe in detail one possible optimization algorithm to train this model.

   (b) (2 points) We could make the neural network "deep" in the sense of stacking sigma functions. Then, a 2-layer network would look like $F(\boldsymbol{X}) = \sigma([\sigma(\boldsymbol{X}\boldsymbol{\beta})]^T\boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ is a matrix of second layer weights. What is the value of added layers?

   (c) (2 points) Explain the problem of "overfitting" which can be especially relevant for deep neural networks, but also other statistical learning models. How would you combat it in our given scenario?

9. (2 points) Your prediction model is probabilistic in the sense that it outputs $F(X) \in (0, 1)$. However, for whatever purposes you are only interested in the classification, i.e. 0 or 1. You can set an arbitrary threshold so that all $F(x) > threshold$ will be counted as 1. How do you choose the optimal threshold?

10. (4 points) You have trained your model to convergence. It is time to evaluate the quality of your predictions.

    (a) (2 points) Name at least three evaluation metrics that could be used in our scenario.

    (b) (2 points) Which would you choose, given that in the data, roughly 3% of all clients have defaulted.

11. (1 point) In practice, it is common to *ensemble* different models together, for example taking the voting average of three models' outputs to predict the final output. Explain why this helps.

12. (3 points) The financial institution is happy with your work and want to try your algorithm on their whole client base, effectively increasing the dataset size to n= 100 000 000. This is too large to fit into the single memory of any machine at your disposal. How do you proceed with training? Name at least three solutions.

---

13. (0 points) Please provide test feedback in the form of rating on a 10 point (10 being most interesting/fun/difficult) scale how:

    (a) (0 points) Interesting (0 - 10)

    (b) (0 points) Fun (0 - 10)

(c) (0 points) Difficult (0 - 10)

the homework (Assignment + Quiz) was.