DS-GA 3001, Introduction to Data Science (PhD)
Prof. Arthur Spirling, Prof. Julia Kempe
Assignment date: Nov 3, 2020

# Homework 2

This homework must be returned to your TA by **9pm Eastern Time, November 14, 2020 (Saturday)**. Late work will incur penalties of the equivalent of one third of a letter grade per day late, and **up to three days only**.

It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be in one of the following formats: (1) A set of answers and a clearly commented `Python` code appendix (use comments to identify code relevant to each answer you produced), (2) A report consisting of clearly marked answers, each accompanied by the relevant code (e.g., a report generated using Python markdown or similar). **In either case, your code must be included in full, such that your understanding of the problems can be assessed.**

---

1. Fairness from the point of view of different stakeholders.

   (a) Which of the following are plausible explanations for the False Positive Rate disparity in COMPAS's risk predictions, discussed in this ProPublica article? For each (A)-(D) below, state whether it's a plausible explanation, and briefly justify your answer (1-2 sentences per item): (A) Northpointe failed to make COMPAS blind to race; (B) COMPAS uses machine learning, which picks up and reproduces historical patterns of injustice; (C) COMPAS uses machine learning, which performs worse for minority groups because less training data is available; (D) Northpointe defines recidivism as "a fingerprintable arrest involving a charge and a filing of any uniform crime reporting (UCR) code", and it may be the case that members of minority groups are arrested at higher rates.

   (b) Consider again the COMPAS investigation by ProPublica, and additionally consider Northpointe's response to ProPublica (you may wish to consult Northpointe's report). For each criterion (A)-(E) below, explain in 1-2 sentences which stakeholders would benefit from a model that optimizes that metric and why. If you believe that a criterion is not reasonable in this case, state so. (A) Accuracy; (B) Positive predictive value; (C) False positive rate; (D) False negative rate; (E) Statistical parity (demographic parity among the individuals receiving any prediction).

2. This exercise relates to the bias-variance trade-off derivation we had in Lab7. Assume the standard statistical learning setup: $\mathcal{X}, \mathcal{Y}, \mathcal{A}$ are input, output and action spaces, and $(X, Y) \sim P$ is the joint data distribution; for this exercise, consider the square loss.

   (a) Define a hypothesis space $\mathcal{F} = \mathcal{A}$ of constant functions. Prove that the Bayes decision function from $\mathcal{F}$ is $a^* = \mathbb{E}Y$. Compute the associated Bayes risk.

(b) Prove that the Bayes decision function (not restricted to $\mathcal{F} = \mathcal{A}$) is $\bar{y}(X) = \mathbb{E}_{Y|X}Y$ (hint: recall from Lab7 that $\mathbb{E}_{(V,W)}g(V,W) = \mathbb{E}_V\mathbb{E}_{W|V}g(V,W)$ for any random variables $V, W$ and any function $g$).

3. This exercise refers to the `question3_student.ipynb` Jupyter notebook; you will prepare data for the modeling problem in the next question. The five files distributed together with this assignment are `rpl17.html`, `rpl18.html`, `rpl19.html`, `rpl20.html`, `rpl21.html`, which contain detailed statistics on soccer matches in the last five seasons of the Russian Premiere League. Your task is to complete the `process_season` function. Please follow the more detailed instructions in the notebook.

4. This exercise refers to the `question4_student.ipynb` notebook. You will build, train, optimize and analyze a binary logistic regression model to predict if a soccer match will end in a home win or not using data produced in Question 3 (or you can use `.npy` files provided with this assignment). The binarization of target values is done for you in the beginning of the notebook (positive class corresponds to a home win). Throughout this question, you should not use `scikit-learn` library.

   (a) The training data is likely unbalanced; upsample the smaller class to achieve a perfectly balanced training dataset (your solution should work both when the smaller class is positive and when it is negative). Since our dataset is still rather small, it is wise to use cross-validation; prepare five different training/validation splits for a 5-fold cross-validation.

   (b) Update the `LogisticRegression` class from Homework 1 to allow for $L2$ regularization (note the argument $C = \lambda^{-1}$, where $\lambda$ is the regularization strength). Recall that the loss function for this model is

$$l(\theta) = -\sum_{i=1}^{n} x_{ij}\left(y_i - \sigma\left(\theta^T \bar{x}_i^T\right)\right) + \lambda\|\theta\|_2^2.$$

   (c) Complete the `validate` function that plots the ROC curve for each of the $K = 5$ classifiers (one per fold). In addition, the plot should contain a pointwise average of those ROC curves and a dashed diagonal from $(0, 0)$ to $(1, 1)$. Finally, compute the mean AUC score. Note that you cannot use `scikit-learn` in this question.

   (d) Fit Logistic Regression models using 5-fold cross-validation for each value of $C$ provided in the notebook. Choose reasonable values for the learning rate $\alpha$ and `max_iter` (recall that logistic loss is convex, so gradient descent is guaranteed to converge given sufficiently small $\alpha$; thus, increasing iterations and reducing $\alpha$ is a safe bet). Apply the `validate` function you prepared in the previous question for each selection of $C$. Display the 8 resulting plots and comment on the optimal choice of $C$. Note that subplot instances are already prepared in the code.

   (e) Complete the `confusion_matrix`, `precision`, `recall`, and `f1_score` functions. Then, use cross-validation to select the appropriate threshold for prediction by optimizing the mean (across folds) $F1$ score. Use $C$ selected in the previous part. Report your findings.

   (f) CSKA will be playing against Rostov in Moscow (home stadium for CSKA) on November 9th. The odds from bookmakers are 1.81 for CSKA and 1.99 for Rostov or a draw (which

means that betting \$1 on CSKA could win you \$1.81 (including your bet of \$1), while betting on Rostov or a draw—\$1.99). The current statistics on both teams are prepared for you in the file `CSKA-Rostov.npy`. Train your best model on all training data, compute the confusion matrix on the test set and assess the expected profit brought by your model under these odds (recall that the positive class, in this case, is the victory of CSKA). Then, compute the prediction and check after Novemeber 9th if it was correct!

(g) What are some disadvantages of predicting the outcome of soccer matches in this way and how the modeling pipeline could be improved?

5. In this problem you will relate Logistic Regression (LR) and Support Vector Machines (SVM) for binary classifiers. We will go from LR to SVM. Recall that the LR computes a probability $p_{LR}(y = 1|x) = \sigma(\omega x+b)$ for each data point $x$. Let us now assume that instead of maximizing MLE to get to the parameter $\omega$, we want to define a classifier from $p_{LR}$ as follows:

$$\hat{y} = \begin{cases} 1 & \text{if} \quad p_{LR}(y = 1|x) \geq p_{LR}(y = 0|x) \\ 0 & \text{otherwise} \end{cases}$$

Now, impose that this classifier, on training data, needs to make the right decision with some *margin of certainty*: either $p_{LR}(y = 1|x) \geq c \cdot p_{LR}(y = 0|x)$ or $p_{LR}(y = 0|x) \geq c \cdot p_{LR}(y = 1|x)$. Among all the solutions for $\omega$ we will chose the one that minimizes its L2-norm. Derive that the classifier is a *hard-margin SVM* and give a choice of $c$ to match to the definition of SVM.

6. SVD - Lily's question:

(a) Explicitly show with a small example (e.g. 3-4 points) that the least squares regression line differs from the first PCA direction.

(b) Further, in 2D, show that PCA minimizes *perpendicular L2-distance* between the first component line and the data points.

7. PCA: Assume your data set consists of $M$ points of dimension $p$ coming from the following distribution: (1) an index $i \in \{1, \ldots, p\}$ is picked uniformly at random. Then (2) the data point is the vector $a\delta_i = (0, \ldots, 0, a, 0, \ldots, 0)$, where $a$ is in the $i$th position and comes from an arbitrary distribution $P(a)$ (let's say it is bounded).

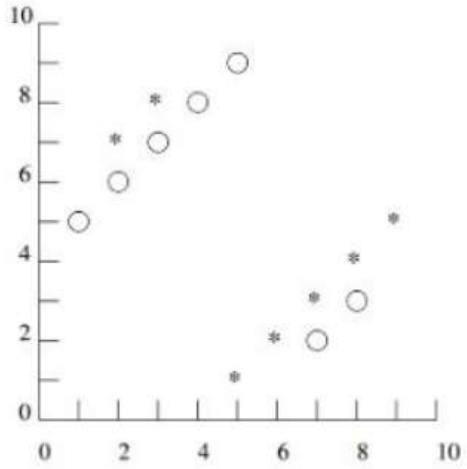
Compute the first principal vector of the design matrix and show that all other eigenvalues are equal (Hint: the covariance matrix is of the form $C_{i,j} = \lambda + \mu\delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise). Discuss whether PCA is a good way to select features for this problem.

8. k-NN:

(a) Let $k$-NN(S) be the $k$ Nearest Neighbor classification algorithm on sample set S, which takes the majority of the closest $k$ points where there are 2 classes (positive and negative).

- Show that if in both 1-NN(S1) and 1-NN(S2) the label of point $x$ is positive, then in 1-NN($S1 \cup S2$) the label of $x$ is positive.
- Show an example such that in both 3-NN(S1) and 3-NN(S2) the label of $x$ is positive, and in 3-NN($S1 \cup S2$) the label of $x$ is negative.

(b) One of the problems with k-nearest neighbor learning is selecting a value for k. Say you are given the following data set. This is a binary classification task in which the instances are described by two real-valued attributes.




- What value of k minimizes training set error for this data set, and what is the resulting training set error? Why is training set error not a reasonable estimate of test set error, especially given this value of k?
- What value of k minimizes the leave-one-out cross-validation error for this data set, and what is the resulting error? Why is cross-validation a better measure of test set performance? (Note: leave-one-out cross validation on n points trains on $n-1$ points and tests on the last point, for each way to remove one point. It hence runs the model $n$ times.)
- Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?
- Sketch the 1-nearest neighbor decision boundary for this dataset.