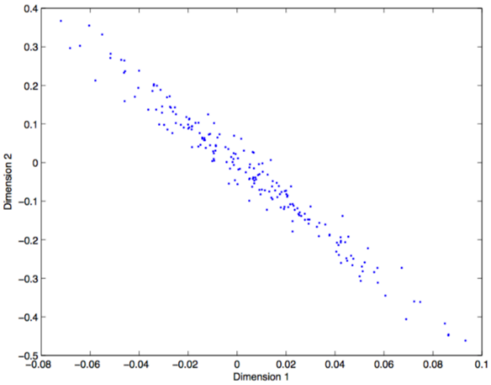
Academic year 2020-2021, January 19, 13h00

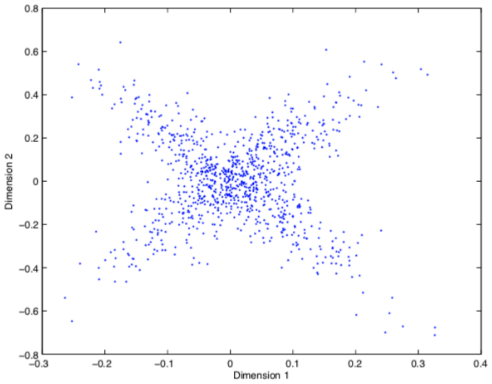
Machine learning

Clearly write your name on every answer sheet, as well as on all question sheets. Answer each question separately, and clearly mention the number of every question next to your answer. If you don't provide an answer for a question, clearly mention the question number and write "No answer". You can answer either in Dutch or English.
No laptops, calculators, PDAs, phones or Internet access is allowed. Hand in all sheets, including question sheets.

1. **Dimensionality reduction - PCA (1 pt)**

   Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principle components on each plot.





2. **Short questions (5 pt)**
   A few sentences should be sufficient to answer each question.

   (a) Describe how counterfactuals can be used to assess algorithmic fairness.

   (b) Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering.

   (c) You are working on a high-dimensional clustering problem, where it is known that the dataset contains many irrelevant and redundant features. Suggest two possible ways of reducing the feature space that will work in this unsupervised setting.

   (d) What will be the different effect on the weights of the regression model when using ridge regression versus LASSO ?

   (e) You design a fully connected neural network architecture where all activations are sigmoids. You initialize the weights with large positive numbers. Is this a good idea? Explain your answer.

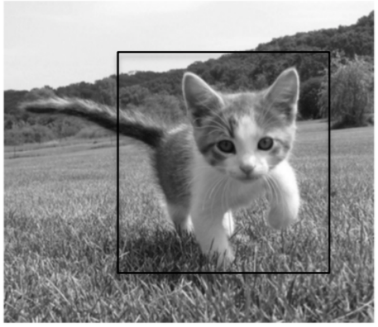
3. **Method design (3 pt)**
A project team is confronted with the binary classification task of detecting credit card fraud in a dataset that contains 50 fraudulent transactions and 5000 normal transactions. The data is high-dimensional, consisting of 1000 features describing the data. They proceed as follows.

First they decide to balance the data by applying SMOTE to upsample the minority class, and obtain a dataset with 5000 fraudulent and 5000 normal transactions. Next, they apply 10-fold crossvalidation, resulting in 10 training and 10 testing folds. For each fold, they compare a KNN classifier, a Random Forest and a linear SVM classifier. As they are afraid that the KNN classifier will not generalize well in high-dimensional spaces, they decide to combine the KNN classifier with PCA, while the SVM and Random Forest classifiers are better suited for high-dimensional data, so they are run on the full (original) feature set. As the PCA is not using any class labels, they run PCA once on the full dataset, keeping the top N principal components that cover 90% of the variance of the data. This reduced dataset is then used as input for the KNN classifier.

As they notice that some methods have hyperparameters (K for KNN and C for the SVM), they conduct a grid search for these parameters. They decide to use a simple way of doing this, which does not require any complex programming: they just train the model for every hyperparameter value on the training part and also test it on the training part within each fold. For every fold, they then choose the hyperparameter value with the best accuracy on the training part of the fold, and evaluate this model on the test part of the fold. In this way, they get for each of the three models 10 accuracy values, one for each of the 10 folds of the crossvalidation. To compare these models in a statistical way, they consider pairwise comparisons between all pairs of models using the K-Fold CV Paired t Test.

Describe the flaws in the above design, explain why you think they are flaws and how they can be mitigated. Your answer should not exceed one page.

4. **Neural networks (3 pt)**
You have a friend who is an animal lover, and he asks you to build a neural network based classifier for object detection of animals in images. For this specific task, your friend is not only interested in classifying every image, but also in localising every object. As an example, consider the picture of a cat below, where next to producing a class label, your network should also predict the location of the cat, using a simple rectangular bounding box.



You are given a large class-balanced dataset of 10,000 images, each annotated with a class label, that represents one of ten possible class labels (cat, dog, cow,...) and a bounding box that represents the location of the animal. Your task is to design a single neural network that can both perform the classification as well as predict the bounding box for unseen images. You can safely assume that each image contains only one animal and all images are of the same size and resolution.

Clearly specify the architecture you propose to use to solve this problem, as well as the loss function and the way you will train the model. Make a drawing of the structure of your architecture, and clearly describe the input, output, and any intermediate layer(s) you would use. Your answer should not exceed one page.

5. **Extensions of the classical supervised learning paradigms (4 pt)**

   (a) Describe how the SVM classifier can be adapted to deal with multi-instance classification problems.

   (b) Two historians approach you for your deep learning expertise. They want to classify images of historical objects into 3 classes depending on the time they were created: antiquity (class 1), middle ages (class 2) and modern era (class 3). Over the last few years, the historians have collected nearly 5,000 hand-labelled RGB images. You start experimenting with convolutional neural networks, but find out you need more training data. After a while, you discover online a very similar network that is already trained on 1,000,000 historical objects from a slightly different time period. You can get access to the parameters of that model, but not the original data set. How will you proceed ?

6. **Bias and fairness (1 pt)**
   You are hired for a new machine learning spin-off company in Gent, that wants to set an example in the field. The CEO asks you to suggest three ways of improving the transparency of their machine learning models. What would you suggest ? Try to be concise and to the point (half a page maximum should be sufficient here).

7. **True or False (3 pt)**
   Are the following statements True or False ? If True, explain in at most two sentences. If False, explain why or give a counterexample in at most two sentences.

   (a) Consider k-fold crossvalidation. With a higher number of folds k, the estimated error will be, on average, higher than with a lower number of folds.

   (b) Consider a regularized machine learning model. More regularization will then lead to a higher bias and lower variance component of the error.

   (c) As the number of data points grows to infinity, the MAP estimate approaches the MLE estimate for all possible priors. In other words, given enough data, the choice of prior is irrelevant.

   (d) Recall is an appropriate performance measure that can be used for a machine learning problem where you have only positive and unlabeled data points.

   (e) In batch normalization, the mean over all features is calculated, which is then used to mean-center all features. This will lead to faster convergence.

   (f) Consider learning a classifier in a situation with 1000 features in total. 50 of them are truly informative about class. Another 50 features are direct copies of the first 50 features. The final 900 features are not informative. Assume there is enough data to reliably assess how useful features are, and the feature selection methods are using good thresholds. Then a wrapper based feature selection method will select a feature subset of 50 features.