Written examination
DIT865/DAT340: Applied Machine Learning, August 22, 2018

Course responsible: Richard Johansson, CSE (+46 31 772 1887)

Allowed accessories: Calculator, written notes (**one** A4 paper)

Please note:

- If there is something you don't understand about a question, please ask the course responsible to clarify when he comes to the exam room (at about 15:00 and 17:00).
- Make sure that your handwriting is legible. You will get no points for unreadable solutions.
- If your solution to a question is incomplete, please turn it in anyway! Every point counts.

Part 1: Basic questions

You need a score of 27 points in this part to receive a passing grade (*G/3*).

Question 1 of 12: Dermatology (8 points)

In dermatology, correctly diagnosing various types of skin conditions can be challenging, since the symptoms of the different conditions can be quite similar. The skin conditions that we consider are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Determining the exact condition can sometimes require a biopsy, which can be expensive.

In order to speed up the process of making a diagnosis, we'd like to develop an automatic classifier. For a large number of patients, we describe their symptoms using 33 attributes. These attributes are a mix of categorical attributes (e.g. the patient's gender, whether or not the skin itches) and numerical attributes (e.g. the patient's age, the size of the affected part). For each of these patients, a diagnosis has been determined: one of the six conditions mentioned above, or a special label *Other* for other types of diagnoses.

(a, 6p) Explain how you would implement a machine learning model that would solve this prediction task. You don't need to show Python code, but please give a description of the system and explain all steps you would carry out when developing it.

(b, 2p) This solution is based on a mix of numerical and categorical features. If we'd like to use an image instead of the features, what type of machine learning model would you suggest?

Question 2 of 12: Consultation (4 points)

Your boss, always ready to jump on the latest bandwagon, decides that your company's existing predictive systems are going to be replaced by machine learning systems. Give an example of a situation where you would recommend the company *not* to switch to a machine learning solution.

Question 3 of 12: Using scikit-learn for a regression task (8 points)

We collect some spreadsheet data and we'd like to build a regression model that learns to predict one of the columns. We use scikit-learn for the implementation, and this is how we build the model:

```
pipeline = make_pipeline(
    DictVectorizer(),
    StandardScaler(),
    SelectKBest(k=100),
    DecisionTreeRegressor()
)
```

Can you describe the purpose of each of the four steps in the pipeline?

Question 4 of 12: Evaluation (8 points)

We have developed a convolutional neural network that classifies images and detects whether they contain a vehicle or not.

(a, 4p) We evaluate the classifier on a test set. Here is the confusion matrix. In the table, + means a positive instance (the image contains a vehicle), and - a negative instance (no vehicle).

		Predicted	
		+	-
Truth	+	20	5
	-	30	945

Compute the accuracy of the classifier, and the precision and recall for detecting vehicles.

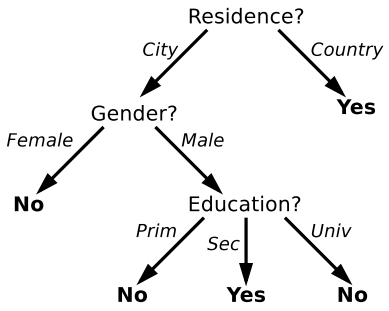
(b, 4p) Compute the accuracy and recall for a *majority-class baseline*. The negative class is the most common in the training set. And why am I not asking you to compute the precision?

Question 5 of 12: Decision tree classifiers (4 points)

The following dataset can be used to train a classifier that determines whether a given person is likely to own a car or not. There are three features: education level (primary, secondary, or university); residence (city or country); gender (female, male).

education	residence	gender	has car?
sec	country	female	yes
univ	country	female	yes
prim	city	male	no
univ	city	male	no
sec	city	female	no
sec	country	male	yes
prim	country	female	yes
univ	country	male	yes
sec	city	male	yes
prim	city	female	no
univ	city	female	no
prim	country	male	yes

Here is a decision tree that was trained from the dataset above.



Explain why the learning algorithm has decided to use the residence feature as the top node.

Question 6 of 12: Music recommendation (8 points)

In a music streaming service, we'd like to build a system that suggests albums by musicians that the user hasn't seen before. This particular streaming service asks the users to rate the albums they have listened to (a numerical grade between 1 and 10).

(a, 4p) Suggest a machine learning method to use for this task. For each user, you have a list of the albums he or she has listened to, as well as the user's rating for those albums.

(b, 4p) Discuss the circumstances where this type of solution works best.

Part 2: Questions for the high grades

DIT865: You need a score of 29 points in this part to receive the grade *VG*.
DAT340: You need a score of 16 for the grade *4*, and 29 for the grade *5*.

Question 7 of 12: Neural network discussion (6 points)

(a, 3p) Why do some people claim that neural network classifiers are less interpretable than a linear model or decision tree? Do you agree? Why does this matter? Explain.

(b, 3p) Why does the selection of a random seed (that is, the initialization of the random number generator) matter more when using neural networks than when using e.g. decision trees? How does this matter for scientific experiments involving neural networks?

Question 8 of 12: Regularizers (6 points)

(c, 2p) What is the purpose of a *regularization* method?

(d, 4p) Describe the implementation of *two* methods of regularization commonly used in neural networks.

Question 9 of 12: Bagging and perceptrons (6 points)

The *perceptron* learning algorithm produces a linear classifier: that is, one that can be expressed in terms of a scoring function

$$\text{score}(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$$

If we apply the *bagging* algorithm in combination with perceptron learning, will this also be a linear classifier? You can assume that the bagging algorithm uses averaging, not voting. Do you think the resulting classifier will work differently (better or worse) than just a plain perceptron?

Question 10 of 12: Support vector classification (8 points)

Support vector classification (SVC) is one of the most popular machine learning models. Training a (linear) SVC model is done by finding the weight vector $\boldsymbol{w}$ that minimizes the function f in the following equation:

$$f(\boldsymbol{w}, \boldsymbol{X}, Y) = \sum_{i=1}^{n} L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + R(\boldsymbol{w})$$

As usual, $\boldsymbol{X}$ is a list of feature vectors $\boldsymbol{x}_i$ of all the instances in the training set and Y the corresponding outputs y_i (each output coded as +1 or -1). λ is a user-defined parameter.

The loss function L is defined

$$L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \max(0, 1 - y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x}_i)).$$

and the regularizer R is

$$\frac{\lambda}{2} \cdot \|\boldsymbol{w}\|^2$$

where λ is a user-defined constant.

Show how these equations can be turned into an actual algorithm (including pseudocode) for training the classifier. The gradient of the loss L with respect to the weight vector $\boldsymbol{w}$ is

$$\nabla L = \begin{cases} -y_i \cdot \boldsymbol{x} & \text{if } y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x}) < 1 \\ (0, \ldots, 0) & \text{otherwise} \end{cases}$$

The gradient of the regularizer R is

$$\nabla R = \lambda \cdot \boldsymbol{w}$$

Question 11 of 12: Ordinal regression (6 points)

An *ordinal regression* task is a prediction problem where we'd like to predict ordered categories. Examples of ordinal regression tasks would be to predict the grade of a student using a non-numerical scale, e.g. the ECTS grade scale of A–F, or a star rating for a movie review.

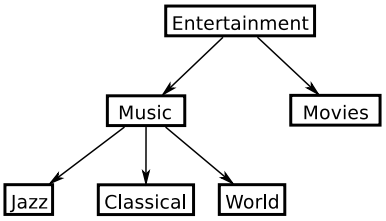
(a, 4p) There are some learning algorithms specifically designed for solving ordinal regression tasks, but let's ignore that for now. Mention *two* different approaches to solving ordinal regression tasks using simpler and more "standard" tools, such as scikit-learn's linear prediction models. Discuss their advantages and drawbacks.

(a, 2p) How would you evaluate a predictor that outputs an ECTS grade?

Question 12 of 12: Hierarchical classification (6 points)

The *Large-scale Hierarchical Text Classification* (LSHTC) Challenge was a series of machine learning competitions held between 2009 and 2014.[1] The focus of these competitions was document classification using a very large set of document categories (roughly 325,000).

The categories were *hierarchically* organized. The figure below shows a tiny fragment of such a hierarchy. You can assume that the hierarchy is tree-structured: that is, that every category (except for the special "root" category) has exactly one parent category.



(a, 2p) How do you think a classifier should be evaluated in this competition?

(b, 4p) Propose *two* different approaches to classifying documents under these circumstances. Discuss their advantages and drawbacks.

[1] `http://lshtc.iit.demokritos.gr/`