1 Specifying ML Tasks

For the following problem descriptions, explain whether the problem is a supervised or an unsupervised problem and whether it is classification, regression, or clustering.

1. We have a set of images, with descriptions of what object is depicted in the images. Given new images, we want to decide what object is on it.
2. Given part of a text, we want to predict the next word.
3. We want to predict the income of graduates of the university. Only few people answer our question, but we know via social networks who they are friends with.
4. We want to identify genres of books given a large collection of them.

2 Generative vs Discriminative Models

Explain the difference between a generative and a discriminative model. Give an example for each of them.

3 Definitions

Give formulas for the following terms. Define all variables you use.

1. Mean squared error
2. Mean absolute error
3. Cross-entropy loss

4 Hyperparameters

What is the difference between hyperparameters and parameters. Give an example of a hyperparameter for a logistic regression model with L2 regularization.

5 Model Evaluation

1. For the confusion matrix in Table 1, calculate precision, recall, F1-score, and accuracy.

Table 1: Confusion Matrix

		Prediction	
		positive	negative
Ground Truth	positive	8	2
	negative	16	974

2. Given the output $\hat{y}$ of a classifier and the ground truth labels y, calculate and draw the ROC curve:

$$\hat{y} = [0.5, 0.35, 0.8, 0.1, 0.2]^T$$
$$y = [1, 0, 1, 0, 1]^T$$

3. Explain when using a ROC curve make sense compared to just using accuracy etc.
4. What is a problem of the ROC curve in heavily imbalanced classification problems?
5. Explain the tradeoff between k-fold crossvalidation and splitting data into a single train and test part.

6 Information Leakage

1. Define the term information leakage with respect to model training.
2. Decide, whether the following scenarios have information leakage. Explain your decision.
   (a) You want to train an object detector on a video of football. To later on estimate the performance of the detector, you split the data randomly into training and test parts. You only use the training data to fit the model.
   (b) You want to classify gene data. Since each instance has 10000 features, you first want to select the top k features to avoid overfitting. You first run a statistical test to select these features, then you split the data into training and test parts.
   (c) You want to classify nodes in a citation graph with respect to the contents of the documents. You augment the nodes with word frequencies, which are used as features for the classifier. Then you split the data.

7 KNN Classifier

Given the data points $[0, 1, 4, 13, 11]$ with labels $[0, 2, 2, 1, 1]$, calculate the predictions of 3-NN classifier (by L1 distance) with majority voting for the points $[3, 10, 6]$.

8 Over- and Underfitting

1. Define overfitting.
2. Define underfitting.
3. Explain the role that regularization plays with respect to over- and underfitting. Give two examples of regularization techniques.
4. Draw a plot that represents an underfitted, overfitted, and well fitted model, respectively.

9 Scikit-Learn

Solve the uploaded Jupyter notebook as an introduction to scikit-learn.