Exam
Machine Learning I: Foundations
Summer Term 2020

Assignment Sheet

First make sure that your exam is complete:

- The assignment sheet (this sheet!) should be made up of 3 pages with assignments 1, . . . , 6.
- The answer sheet should be made up of 9 pages. 1 cover sheet, 1 page per assignment 1, . . . , 6, and 2 extra sheets.

The point total of this exam is 100.

If you find it helpful, you may remove the paper-clip from the answer sheet. However, at the end of the exam, please bring the pages back in the order as you received them. If you used extra pages you may insert them at the appropriate place.

Exam ID:

Assignment 1 (True or False)

$10 * 1.5 = 15$ points

For each of the following sets of statements exactly one of the statements is true. Determine the true statement. For each correct answer you receive 1.5 points. For each incorrect answer you receive -1 points. Not giving an answer awards 0 points. The total for this assignment cannot drop below 0 points.

a) Random forests are a machine learning model in which ...

1. a decision tree is trained for each label and a data point is labeled by all the trees.
2. a forest is trained and during evaluation a random walk through this forest is generated and its path nodes are used as the label.
3. multiple decision trees are trained, which vote during evaluation.

b) K-means is ...

1. an unsupervised algorithm for feature extraction.
2. an unsupervised algorithm for determining clusters.
3. a supervised algorithm which evaluates a label by taking the mean of k ML models.

c) The non-linearity of neural networks originates from ...

1. the stacking of multiple layers.
2. its activation function.
3. its use of kernels.

d) In neural networks the choice of activation function is usually ...

1. not limited by any factor.
2. limited to differentiable functions.
3. limited to functions which are differentiable almost everywhere.

e) Least squares regression ...

1. can be trained by gradient descent.
2. is an algorithm which tries to find a minimum square which encapsulates all data points.
3. cannot evaluate the label of a data point with $\mathcal{O}(d)$ computations.

f) Ridge regression ...

1. can be trained by inverting one matrix in addition to a few additions and multiplications.
2. is much more efficient than least squares regression.
3. cannot evaluate the label of a data point with $\mathcal{O}(d)$ computations.

g) In the k-nearest-neighbour algorithm ...

1. most computations are done during training.
2. training involves approximately as many computations as testing.
3. most computations are done during testing.

h) In machine learning, kernels are considered ...

1. functions that efficiently compute inner products after a mapping.
2. non-linear machine learning algorithms.
3. important for any machine learning algorithm.

i) A machine learning model is usually ...

1. a beautiful woman posing in front of machine learning.
2. not worth using in practice due to its complexity.
3. a parametrized function used for prediction.

j) In general a machine learning algorithm ...

1. can be kernelized.
2. can be used for any problem.
3. can be used to analyze data.

Level of expectation: Your answers should be of the form a)1, b)2, c)3. Make sure to put your answer on the answer sheet, otherwise your answer will not be graded.

Assignment 2 (Regression)

5 + 5 = 10 points

Let $X \in \mathbb{R}^{d \times n}$ be the data matrix, $\mathbf{y} \in \mathbb{R}^n$ be the label vector, $C \in \mathbb{R}$ be the regularization parameter, and $\mathbf{s} \in \mathbb{R}^d$. Consider the following regressor.

$$\mathbf{w}_{TRR} := \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w}\|^2 + C \left\|\mathbf{y} - X^T\mathbf{w}\right\|^2 - \mathbf{s}^T\mathbf{w}$$

a) Derive a closed form solution for $\mathbf{w}_{TRR}$.

b) The above equation is a slight variation on ridge regression. State the difference between ridge regression and the above regression, including a geometric interpretation.

Level of expectation: In item b) your geometric interpretation might be at odds with other terms of the objective. Resolving these issues grants more points.

Assignment 3 (Principal Component Analysis)

$5 * 3 = 15$ points

Let $X \in \mathbb{R}^{2\times n}$ be the data matrix, $S_n = (X - \hat{\mu})(X - \hat{\mu})^T$ be the scatter matrix, and

$$S_n = \begin{bmatrix} 3 & -4 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} 0.63 & 0 \\ 0 & 2.74 \end{bmatrix} \begin{bmatrix} -\frac{3}{7} & -\frac{4}{7} \\ -\frac{4}{7} & -\frac{3}{7} \end{bmatrix} \text{ be its diagonalization.}$$

a) What is the first principle component of the data according to PCA?

b) Project the point $\mathbf{x} = \begin{bmatrix} 2 & -4 \end{bmatrix}^T$ using PCA with $k = 1$ into the coordinate system with the basis made up of the first principle component.

c) What happens to data points if we project them using PCA as in c) but with $k = 2$? Illustrate this using a sketch.

d) What could the process from d) be used for?

e) Assume that the S_n above is not centered and we do not have access to the data points. Can we use the same trick we used to center the kernel matrix to center S_n? Explain your answer.

Level of expectation: In item a) your answer should just be the principal component. In item b) you should include the derivation of the projection.

Assignment 4 (Support Vector Machines)

$2 + 2 + 5 + 6 = 15$ points

Consider the following binary dataset. Each element is a tuple of the data point and its label, i.e. $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i, y_i \in \mathbb{R}$.

$$D := \left\{ \left(\left[-3\right], 1 \right), \left(\left[-2\right], 1 \right), \left(\left[-1\right], 1 \right), \left(\left[0\right], -1 \right), \left(\left[1\right], -1 \right), \left(\left[2\right], -1 \right) \right\}.$$

a) State the objective function of the soft-margin SVM, using the hinge-loss.

b) State what property has to hold for a vector to be a support vector.

c) Consider the case where $C = 0$, state an optimal $\mathbf{w}$ and b and state the number of support vectors.

d) Consider the case where $C \to \infty$, state an optimal $\mathbf{w}$ and b and state the number of support vectors.

Level of expectation: In item a) you may consider the constrained or unconstrained version. For items c) and d), argue why your given amount of support vectors is correct. For items c) and d) you must not include the derivation of the optimal $\mathbf{w}$ and b, however if your found $\mathbf{w}$ and b are incorrect this might give you partial points.

Assignment 5 (K Nearest-Neighbor)

$5 + 5 + 5 + 5 = 20$ points

```
1  function knn_train(parameter k, data X = {x_1, ..., x_n}, labels y)
2  function knn_evaluate(input x)
3  d.append((distance(x_i, x),y_i))
4  initialize d as an empty list
5  sort d ascending by the first item
6  return the vote of d[0:k]
7  for i = 1 : n do
8  for i = 1 : k do
9  store y
10 store X
11 store k
12 end for
13 end function
```

a) Use the above lines to code the k nearest neighbor algorithm, this includes the training as well as the evaluation of k nearest neighbor. You may use lines twice. Some lines are traps, i.e. they are not needed.

b) Prove that KNN can be kernelized.

c) How does changing k relate to underfitting and overfitting. Explain your answer.

d) KNN is sometimes called an unusual ML algorithm as evaluation computationally takes more time than training. Explain why this is unusual.

Level of expectation: For item a), you should only write down the sequence of numbers, i.e. 6,3,2,4,5,1, there is a line to separate functions. The **store** keyword allows variables to be used later on in other functions. For item c) you may assume that you know the perfect k, named k^*. For item d) you should especially talk about algorithms other than KNN.

Assignment 6 (Kernels)

$5 + 5 + 10 + 5 = 25$ points

Let $c, d, p \in \mathbb{N}$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, and let k_1 and k_2 be kernels.

a) Prove that $k(\mathbf{x}, \mathbf{x}') := k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$ is a kernel.

b) Prove that the polynomial kernel, $k(\mathbf{x}, \mathbf{x}') := (\mathbf{x}^T\mathbf{x}' + c)^p$ is a kernel.

c) Consider the feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$ for the polynomial kernel, such that $(\mathbf{x}^T\mathbf{x}' + c)^p = \langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$. Explicitly derive the formula for m involving d and p.

d) Explain why knowing the dimensionality of a feature map for a kernel might be helpful.

Level of expectation: In item b) you may use all kernels and kernel theorems stated during the lecture and the exercises, except for the statement you are proving, however you have to separately state them and then refer to them. You may also use a) to solve b). In item c) you should also explain your derivation based on a) and b). In item d) you may consider the case of the kernelized SVM instead of the general case.