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, . . . , 7.
- The answer sheet should be made up of 10 pages. 1 cover sheet, 1 page per assignment 1, . . . , 7, 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)

$2 + 2 + 2 + 2 + 2 = 10$ 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 2 points. For each incorrect answer you receive -2 points. Not giving an answer awards 0 points. The total for this assignment cannot drop below 0 points.

a) Backpropagation is an algorithm whose primary and direct purpose is to ...

1. optimize the parameters of neural networks.
2. efficiently compute the gradients occurring in neural network training.
3. compute the inverse of a neural network.

b) Support vectors are ...

1. points added to the data to support and stabilize optimization.
2. the directions along which SVMs are supported.
3. the points which define the separating hyperplane.

c) Gradient descent is an optimization method that ...

1. can find the global minimum of any function.
2. is uniquely used to optimize neural networks.
3. can find the global minimum of a function under some assumptions on the function.

d) Linear machine learning algorithms ...

1. can be turned into non-linear algorithms by using an upstream mapping (*in* $g(f(x))$ f *is an upstream mapping to* g).
2. can be turned into non-linear algorithms only if they can be kernelized.
3. cannot be turned into non-linear algorithms.

e) Kernels are functions that ...

1. efficiently compute inner products.
2. compute any distance measure.
3. find optimal solutions for SVMs.

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 (Convolutional Neural Networks)

$5 + 4 + 2 + 4 = 15$ points

The following equation can be used to calculate the output size after a convolutional layer.

$$o = \frac{i - k + 2p}{s} + 1,$$

where i is the input size, o is the output size, k is the kernel size, p is the padding, and s is the stride.

a) Put this equation into the context of convolutional neural networks, by explaining the variables i, o, k, p, and s *in more detail than this assingment.*

b) When talking about a convolutional layer the quantities c, the number of channels, and f the number of filters, are usually also important. However, they dont appear in the above formula. Why is that? What changes in a convolutional layer if you change c or f?

c) Give one example of what the above formula could be used for, or why it is important.

d) Consider the following, we want to change a convolutional layer to not use only one kernel size k, but two different sizes, because we expect the spatial dimension of features to vary. What problems will occur and how can they be resolved? *In this assignment consider two kernels like* 3×3 *and* 5×5*, not a non-square kernel like* 3×5.

Level of expectation: The formula given is a simplification. The complete formula would involve rounding, however this should not make a difference in working on the assignments.

Assignment 3 (Regression)

$2 + 5 + 3 = 10$ 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.

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

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

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

c) The above regressor is actually equivalent to ridge regression, i.e., for all $\mathbf{s} \in \mathbb{R}^d$, it is possible to formulate ridge regression such that $\mathbf{w}_{CRR} = \mathbf{w}_{RR}$. Explain how this is possible. Hint: use the result from b).

Assignment 4 (Stochastic Gradient Descent)

4 + 3 + 3 = 10 points

1 $\theta_t \leftarrow \theta_{t-1} - \lambda_t \nabla_{\theta_{t-1}} f(\theta_{t-1})$
2 **for** t = T : 1 **do**
3 **function** GD(parameter T, function f, learning rate λ_t)
4 **return** θ_T
5 $\theta_t \leftarrow \theta_{t-1} + \lambda_t \nabla^2_{\theta_{t-1}} f(\theta_{t-1})\theta_{t-1}$
6 **end for**
7 **for** t = 1 : T **do**
8 $\theta_t \leftarrow \theta_{t-1} + \lambda_t \nabla_{\theta_{t-1}} f(\theta_{t-1})$
9 initialize θ_0

a) Use the above lines to code gradient descent. You may use lines twice. Some lines are traps, i.e. they are not needed.

b) We used a variation on your code in class to optimize the SVM. This involved a batch size and data points. Elaborate on how your code has to change to incorporate a batch of data points.

c) When using SGD to optimize the soft-margin SVM, some renormalization involving the batch size and number of data points is necessary. State this normalization constant and explain why it is necessary.

Level of expectation: For item a), you should only write down the sequence of numbers, i.e. 6,3,2,4,5,1. For item b) you need not write code, however you can if it helps you make your point.

Assignment 5 (k-Means Algorithm)

5 + 10 = 15 points

1 **end for**
2 **return** cluster centers $c_1, \ldots, c_k$
3 **function** KMEANS(parameter k, inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
4 **for** $i = 1 : n$ **do**
5 $y_i := \arg\min_{j=1,\ldots,k} \|\mathbf{x}_i - \mathbf{c}_j\|^2$
6 initialize cluster centers $\mathbf{c}_1, \ldots, \mathbf{c}_k$
7 **repeat**
8 **for** $i, j = 1 : n$ **do**
9 **until** convergence criterion is met
10 $\mathbf{c}_j := \text{mean}(\{\mathbf{x}_i : y_i = j\})$
11 **for** $j = 1 : k$ **do**

a) Use the above lines to code k-means. You may use lines twice. Some lines are traps, i.e. they are not needed.

b) Kernelization means replacing inner products $\mathbf{x}_i^T \mathbf{x}_j$ with the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. However, inner products do not appear in the above k-Means code. How then can k-Means be kernelized?

Level of expectation: For item a), you should only write down the sequence of numbers, i.e. 6,3,2,4,5,1. For item b) justify your answer by explicitly deriving stated equations.

Assignment 6 (Kernels)

$6 + 10 + 4 = 20$ points

Let $A \in \mathbb{R}^{d \times d}$ be a symmetric and positive semi-definite matrix and let $\mathbf{s} \in \mathbb{R}^d$ be a vector.

a) Prove that $k(\mathbf{x}_i, \mathbf{x}_j) := \mathbf{x}_i^T A \mathbf{x}_j$ is a kernel.

b) Prove that $\left(\mathbf{s}^T \mathbf{x}_i \mathbf{x}_j^T \mathbf{s}\right)\left(\mathbf{x}_i^T A \mathbf{x}_j + \mathbf{x}_i^T \mathbf{x}_j + \mathbf{s}^T \mathbf{s}\right)$ is a kernel.

c) Let $k_1(\mathbf{x}_i, \mathbf{x}_j)$ and $k_2(\mathbf{x}_i, \mathbf{x}_j)$ be kernels. Prove that the following statement is **false**: $k(\mathbf{x}_i, \mathbf{x}_j) := \alpha k_1(\mathbf{x}_i, \mathbf{x}_j) + \beta k_2(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel for any $\alpha, \beta \in \mathbb{R}$.

Level of expectation: In item b) you may use all kernels and kernel theorems stated during the lecture and the exercises, however you have to separately state them and then refer to them. *You may also use a) to solve b).*

Assignment 7 (Support Vector Machines)

$4 + 10 + 4 + 2 = 20$ points

Consider the following data points

$$D := \left\{ \left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 2 \\ 0 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right) \right\}$$

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

b) Optimize the objective from a) using gradient descent. Use the following optimization or hyperparameters

$$C = \frac{1}{2}, \lambda := 1, \mathbf{w}_0 := \begin{bmatrix} 0 \\ 0 \end{bmatrix}, b_0 := 0.$$

Gradient descent will converge after a few (≤ 5) iterations, so T is not relevant.

c) Determine two different points on the separating hyperplane for your solution $\mathbf{w}^*$ and b^* from b).

d) Determine the predicted label of the following point using your solution $\mathbf{w}^*$ and b^* from b)

$$\mathbf{x} := \begin{bmatrix} \frac{1}{2} \\ 2 \end{bmatrix}.$$

Level of expectation: In item a) you should only state an equation. In item b) you should execute gradient descent while explaining the steps you are taking in sufficient detail, to be able to follow along. In item b) state the optimal $\mathbf{w}^*$ and b^*. Note for item b) sometimes the regularization constant is itself normalized as $\frac{C}{n}$, however in this lecture we only use C. In item c) you should state two points on the separating hyperplane and explain how they were derived. In item d) you should determine the label of $\mathbf{x}$ and explain how it was derived. If you are unable to solve item b) use $\mathbf{w}^* = \begin{bmatrix} -\frac{3}{2} & 0 \end{bmatrix}^T$ and $b^* = 2$ to solve items c) and d).