Exam Preparation Sheet

In the following, we will list possible exam exercises which are classified roughy into three aspects/difficulties:

a) Exercises that can be solved from either common sense or information that was in the script. Usually these exercises discuss advantages and disadvantages of certain models, or ask you to predict the outcome for an already learned model, i.e. just applying $\hat{y}$.

b) The core exercises containing most of the computations that have to be done for many machine learning models and associated learning algorithms.

c) Exercises that go beyond the exercises in b). These exercises may also ask you to connect your knowledge from two areas such as hyperparameter optimization for a certain model, explaining the more complex models and the likes.

For whatever happens during the exam, it makes most sense to first scan the exercises and solve the ones where you know how to solve them or at least have an idea how to solve them. Never let yourself get stuck and waste too much time on a certain exercise!! Also keep in mind that the b) exercises yield most points!

Regression

a) Questions

1) A website collects DVD ratings and then uses them to recommend users a DVD. Given are the ratings of two users among all items (1 Star is the worst rating, 5 the best):

Index	User	DVD	Rating
1	A	*The Big Lebowski*	5 Stars
2	A	*Brazil*	1 Stars
3	A	*Titanic*	2 Stars
4	B	*Brazil*	3 Stars
5	B	*The Godfather*	5 Stars
6	B	*Toy Story*	4 Stars

Three different regression models $\hat{r}_1, \hat{r}_2, \hat{r}_3$ are learned and make the following predictions:

Index	$\hat{r}_1$	$\hat{r}_2$	$\hat{r}_3$
1	4.8	3.8	4.9
2	2.4	1.5	1.3
3	2.2	1.5	2.1
4	3.2	3.1	2.9
5	4.7	4.4	4.2
6	4.1	3.9	4.2

Estimate for every regression model the mean squared error and the mean absolute error, which is given by

$$\text{MAE}(\hat{y}, \mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} |y_n - \hat{y}(x_n)|$$

in comparison to the true ratings.

2) What are the general differences betweeen MSE (Mean Squared Error) and MAE (Mean Absolute Error)?

3) Why do we prefer to use the RMSE (Root Mean Squared Error) instead of the MSE?

4) For the following data:

x_1	x_2	y
1	3	6
0	1	6
2	2	3
5	1	-4

a linear regression model given by the parameters

$$\beta = \begin{pmatrix} 4 & -2 & 1 \end{pmatrix}$$

has been learned. Compute its error as well as its AIC by using the negative RSS as logarithm of the likelihood.

$$\log(\mathcal{L}) = -\sum_{i=1}^{N}(y_i - \hat{y}(x_i))^2$$

5) What are the positive, what are the negative aspects of using a nearest neighbor regression?

6) Discuss the computational impact of using nearest neighbor regression, what happens if the training data is large? How can you speed up nearest neighbor computations?

b) Questions

1) Given are the data instances of the example from the lecture (gas consumption):

$$\mathcal{D} = \{(2, 7), (6, 4), (8, 3)\}$$

Estimate the target $\hat{y}$ for $x = 10$ using the method of least squares. The true value is $y = 1$. Estimate the error.

2) Given is following data:

x_1	x_2	y
1	2	0
-1	3	-3
-1	2	-2
1	4	-2
3	1	3

Learn a linear regression by estimating its parameters using normal equations (i.e. the closed form solution)! Do not forget to include the bias term!

3) For the following data:

x_1	x_2	y
1	3	6
0	1	2
2	2	7
5	1	11

a linear regression model given by the parameters

$$\beta = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix}$$

has been learned. Perform a backward search on the employed variables and compare the three resulting models to the *full* model. Which one do you choose in the end? *Hint:* Use Cramer's rule for inversion of 2×2 matrices:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

4) For the following data:

x_1	x_2	y
1	2	2
4	-2	-3.3
2	3	3.5
-5	-1	-2.8
-1	2	1.8

a linear regression model given by the parameters

$$\beta^\top = \begin{pmatrix} -1 & 0.3 & 1.4 \end{pmatrix}$$

has been learned. Perform a (two sided) hypothesis test on all of the parameters, where

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

in order to determine which parameters are significant. What are the resulting standardized coefficients? Use a significance level of $\alpha = 0.05$. Which variables are significant?

5) Given some data, learn a regression tree using the RMSE as decision criterion for which split is the best.

6) Given is the following data

x	y
1	3
2	5
3	7
4	9
5	11

Compute the prediction for $x = 2.75$ and $x = 4.25$ using a K-Nearest Neighbor with $K = 2$.

7) For the data in 6) learn a Kernel Regression using

$$K(x, x_0) = D(\frac{|x - x_0|}{\lambda}$$

and

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & t < 1 \\ 0 & \text{else} \end{cases}$$

and predict again for $x = 2.75$ and $x = 4.25$ using $\lambda = 1$.

c) Questions

1) In the lecture it was shown that for the simple linear regression

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

minimize the residual sums of squares (RSS).

Proove these equations by computing the partial derivatives of the loss function.

Justify that the given solutions are indeed a global minimizer for the loss.

2) A weakness of using Nearest Neighbor Regression is that it is a locally constant classifier. How can this issue be overcome?

3) Another weakness of Nearest Neighbor is that the optimal number of neighbors K to take into consideration is unknown. How can you optimize K?

4) Which other regression models different to linear regression do you know? How do they scale in terms of the features? How do they behave when learning them?

Optimization

a) Questions

1) What are the advantages and disadvantages of using Gradient Descent to minimize a loss function?

2) What is meant by the term overfitting and how it comes to pass? How can you recognize that a model is overfitted?

3) How can you prevent overfitting?

4) Describe the main difference between regular gradient descent and coordinate descent!

5) Describe the main differences between gradient descent and Newton methods. Why do we not use Newton methods that often?

b) Questions

1) Apply gradient descent on the function $f(x) = \frac{1}{4}x^4 + \frac{1}{3}x^3 - \frac{1}{2}x^2$ under the following configurations:

   a) Use step length $\alpha = 0.3$ and starting point $x_0 = -1$ and show the first four iterations. What is your minimum?

   b) Use step length $\alpha = 2$ and starting point $x_0 = -1$ and show the first four iterations. What has happened and why?

   c) Use step length $\alpha = 0.3$ and starting point $x_0 = 0$ and show the first two iterations. What has happened and why?

   Do the same again with $\alpha = 0.8$ and starting point $x_0 = 0.5$ and show the first four iterations. Where is your minimum now?

2) Ridge Regression learns model parameters β by minimizing the following objective function:

$$f(\beta, \lambda) = \sum_{i=1}^{N} (y_i - \beta^\top x_i)^2 + \lambda \|\beta\|_2^2 \qquad \lambda > 0$$

   Learn two regression models using the **closed form solution** for $\lambda = 2$ and $\lambda = 5$ for the following data

x	y
2	0
4	-2
-2	4
-1	3

   Which model performs better for $x = 5$, where $y = -3$ is the ground truth?

2) Stochastic Gradient Descent (SGD) works very similar to normal gradient descent, the key difference is that only one data instance is used per update, i.e. the (regression) loss function for a single instance resolves to:

$$\mathcal{L}(y, \hat{y}(x)) = \frac{1}{2}(\hat{y}(x) - y)^2$$

For a polynomial regression of order two, i.e.

$$\hat{y}(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{l=1}^{p} \sum_{j=l}^{p} \beta_{lj} x_l x_j$$

Compute the update equations of a stochastic coordinate descent for all parameters β_0,β_i and β_{lj} for the single instance loss!

c) Questions

1) What are the differences between using the ℓ_1 and the ℓ_2 regularization? Make a plot how both regularizations work! Which regularization scheme is prefered when?

2) Why do we not simply learn λ using gradient descent, i.e. by deriving the overall objective function with respect to it? Explain what happens if we would.

3) We want to apply a polynomial regression for a given problem. How do we optimize the degree p of the polynomial?

Classification

a) Questions

1) Describe the main difference between classification and regression problems!

2) Describe the differences between the 1-vs-rest and 1-vs-last approaches for multiclass classification. What are the positive, what are the negative aspects of each approach?

3) For the following data containg three labels:

x_1	x_2	y
1	2	1
2	2	1
-1	2	2
-2	1	2
3	-6	3
2	-1	3

three binary 1-vs-rest classifiers have been learned. The resulting model parameters are

$$\beta_1 = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix} \qquad \beta_2 = \begin{pmatrix} 0 & -3 & 2 \end{pmatrix} \qquad \beta_3 = \begin{pmatrix} 1 & 1 & -4 \end{pmatrix} \tag{1}$$

Compute the predictions for all the data points and assign labels accordingly. What is the final misclassification rate?

4) Scientists compared the earth of Iowa which contains a specific bacterium (class 1) with other earth that does not contain it (class 2). They observed the variables x_1 (pH value) and x_2 (nitrogen content). The number of instances pro class, the mean of the vectors and the covariance matrix for both kind of earths is given as follows:

$$\begin{aligned} n_1 = 13, \qquad & n_2 = 10 \\ \mu_1 = \begin{pmatrix} 7.8 \\ 43 \end{pmatrix}, \qquad & \mu_2 = \begin{pmatrix} 5.9 \\ 18.8 \end{pmatrix} \\ \Sigma_1 = \begin{pmatrix} 0.5 & 6 \\ 6 & 140.2 \end{pmatrix}, \quad & \Sigma_2 = \begin{pmatrix} 0.1 & 0.17 \\ 0.17 & 20.2 \end{pmatrix} \end{aligned}$$

Estimate the discriminant functions for both classes. Then, assign the observation $x = \begin{pmatrix} 6 & 52.5 \end{pmatrix}^T$ to one of the both classes.

5) What are the differences of LDA and QDA?

6) How can we prevent a decision tree from overfitting the training data? Explain all regularization schemes.

7) How are hyperplanes defined and how can they be used for classification?

8) Given is the data

x_1	x_2	y
2	2	1
1	3	1
-2	-11	2
-1	-2	2

and a hyperplane given by $\beta_0 = 0.5$ and $\beta = \begin{pmatrix} -1 & -1 \end{pmatrix}$. Draw the points and the hyperplane. Does it seperate the data?

b) Questions

1) A logistic regression is a binary classifier, where:

$$\hat{y}(x;\beta) = \begin{cases} 1 & \text{if} \quad \sigma(\beta^\top x) > 0.5 \\ 0 & \text{if} \quad \sigma(\beta^\top x) \leq 0.5 \end{cases} \tag{2}$$

for $\sigma(t)$ being the sigmoid function:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{3}$$

For the data:

x_1	x_2	y
1	1	1
2	-1	1
-1	-3	0
-3	0	0

perform **one iteration of gradient ascent**! Initialize all your parameters to zero and use a step size $\alpha = 0.5$!

Perform **one iteration of Newton**! Again, initialize all your parameters to zero and use a step size $\alpha = 0.5$!

2) Learn all parameters for a Linear Discriminant Analysis for the following data:

x_1	x_2	y
1	2	1
2	2	1
-2	1	2
3	-6	2
2	-1	2

Assign a prediction to the instance $x = \begin{pmatrix} 1 & 1 \end{pmatrix}$!

3) Given some data, learn a decision tree using the accuracy as decision criterion for which split is the best.

4) Given is the data

x_1	x_2	y
2	2	1
1	3	1
-2	-11	2
-1	-2	2

perform two epochs of the perceptron learning algorithm using $\alpha = 0.5$!

4) Given is the data (binary features)

x_1	x_2	x_3	y
1	1	0	1
1	0	1	0
0	0	1	1
0	0	0	0
1	1	1	1
1	0	0	0
0	1	0	0

learn all conditionals needed for Naive Bayes!

What is the prediction for $x = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$?

c) Questions

1) Show, why the decision boundary of LDA is linear if we use the same covariance matrix for all classes for $K = 2$ classes!

2) Why is the accuracy not a good proxy for choosing the splits when learning a decision tree? What is used instead, and why does it work better?

3) What happens if we apply the perceptron algorithm for data that is not linearly seperable? How do we overcome this issue with Support vector machines?

4) Given is the data

x_1	x_2	y
1	1	1
-1	1	2
1	-1	2
-1	-1	1

define a mapping $\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ such that the data is linearly seperable in $\mathbb{R}^3$!

5) Show how the Naive Bayes classifier is derived!

Clustering

a) Questions

1) What are the differences between clustering and classification?

2) Describe the advantages and disadvantages of K-Means!

3) How can K-Means computation be sped up? This works very similar to K Nearest Neighbors!

4) What is the difference of K-Means to K-Medoids?

b) Questions

1) Given is the following data:

x_1	x_2
1	2
2	2
-2	1
3	-6
2	-1

Assume the first cluster center is $\mu_1 = \begin{pmatrix} 3 & -6 \end{pmatrix}$. Compute the initial cluster centers for K-Means starting with $K = 2$ and $K = 3$ and using the Euclidean Distance.

2) Perform two iterations of K-Means with $K = 2$ using Euclidean distance.

c) Questions

1) How can we optimize the number of clusters K which we are estimating?

2) K-Means is very dependent on the initialization. How do we overcome this issue?

3) Assume we run K-Means with the Tschebycheff Norm ($p = \infty$). What preprocessing do we have to perform on the data in order for to get reasonable results?

4) Which more complex methods for clustering do you know? How are they different from K-Means?

Miscellaneous

a) Questions

1) What are the d_p metrics for $\mathbb{R}^n$?

2) What is the d_p metric for $p = \infty$?

3) Which is the most commonly used d_p metric?

4) Which distance function do you know for strings, i.e. character sequences?

b) Questions

1) For the data

x_1	x_2
1	2
2	2
-2	1
3	-6
2	-1

compute the distance matrix using the Euclidean distance.

2) Compute the edit distance of the strings STONE and HOME!