Exam protocol Data-Driven Algorithms for Vehicle Technologies 06.08.2024

Time:	1.5 hours, very generous, can easily go through the exam a couple of times; was completely done after about 50 minutes
Questions:	Multiple choice, between one and multiple right answers; 39 questions, between 1 and 2 points per question; all in English
Suggestions:	In general quite an easy exam; slides and code equally important, with my personal bias a little towards the code. My suggestion would be to read through the slides and then do the code exercises as requested in the slides once. Write down how models are initialized and parametrized in the code, how model parameters influence charts. Non-programmable calculator allowed but not necessary.
About this doc:	➡No guarantee on the correctness of these questions and answers. Some pictures taken from the slides/code exercises, some not found in the slides (e.g. Q24) but are similar to the ones in the exam. ➡Questions here are not necessarily in order, but in the exam were given in the order of how the topics appear in the script. ➡Questions marked with a <?> are probably not very correct in their question and answer but should still give you a good overview of what was asked. ➡<…> means there were more answers but I did not remember them

Please comment (or create a Pull Request for the markdown file in GitHub) if something is wrong/missing 🙂

Download latest version here:
https://github.com/burner-filehosting/filesharing

1

{"route":"Karlsruhe-Berlin", "length":500, "turns":[{"name":"Street1", "distance":10, "direction":"north"}, {“name”:”Street2”, ”distance”:3, ”direction”:”northeast”}]}

- Represents a JSON file
- Represents a CSV file
- Represents none of the above

2

What does the following code do?

```
(1) import numpy as np
(2) import matplotlib.pyplot as plt

(3) x = np.array([10.4,22,4.9,16,9.1,11,3.5])
(4) y = np.array([1,2,3])
(5) plt.plot(x,y)
```

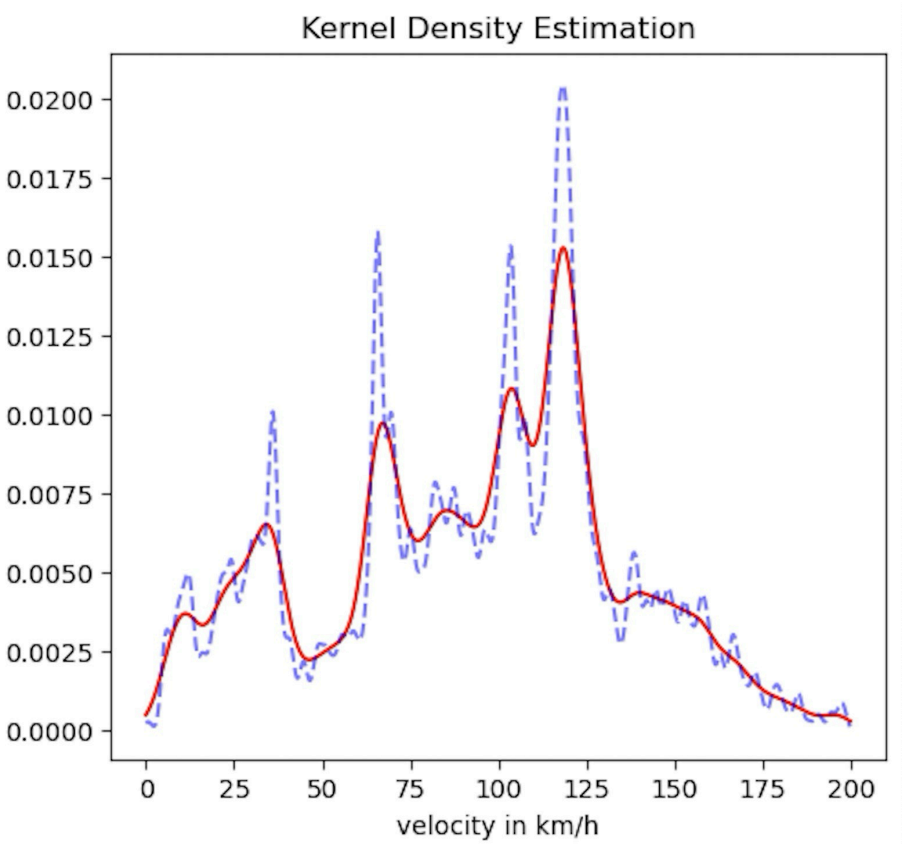
- This will print a plot of y relative to x
- This will give an error in line 3 because the array is not initialized correctly
- This will give an error in line 5 because it cannot be plotted
- This code will run but not output anything

3

What is the difference between a Probability Density Function (PDF) and a Probability Mass Function (PMF)?

- The PDF is used for displaying the relative probability of a continuous random variable at a certain value, while the PMF is used for displaying the probability of a discrete random variable
- The PDF is used for displaying the probability of a continuous random variable at a certain value, while the PMF is used for displaying the probability of a discrete random variable
- PMF and PDF are synonyms and mean the same thing
- …

4



- The red curve represents the model with higher bandwidth
- The blue curve represents the model with higher bandwidth
- The bandwidth will not make a difference for the shown curves

5

You are fitting a model for which you know the function of how parameters and output correlate, but not the parameters.

This model is a

- black box
- gray box
- white box

6

For the least squares regression, the following function is used as a cost function

$$min_{\hat{p}}(y_i - f(x_i\hat{p}))^2$$

- The functions is wrong, as it does not minimize the sum of the residuals
- The functions is correct, as it minimizes the individual squared errors
- The functions is wrong, as it should minimize $f()$
- …

7

What does the following code do?

```
*all variables not initialized in the following code are known*
kmeans = KMeans(n_clusters=2)
kmeans.fit(data)
features = list(zip(v_lim_k_01, u_k_01, kappa_k_01, alpha_k_01, lambda_k_01, kappa_k_01))

inertias = []
for i in range(1,10):
        kmeans = KMeans(n_clusters=i)
        kmeans.fit(features)
        inertias.append(kmeans.inertia_)

plt.figure()
plt.plot(range(1,10), inertias, marker='o')
plt.show()

```

- The code will plot data as estimated by a Support Vector Regression
- The code will plot the “elbow plot”
- …

8

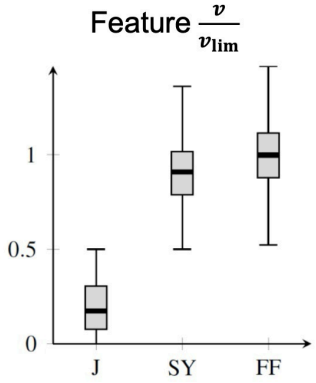
What does the following code do?

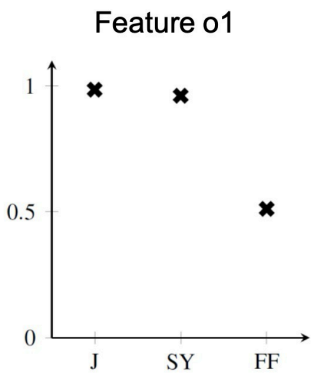
```
i_jammed = [i for i, value in enumerate(traffic_labels) if value == 1]
i_synch = [i for i, value in enumerate(traffic_labels) if value == 2]
i_free = [i for i, value in enumerate(traffic_labels) if value == 3]

data_1 = v_ego[i_jammed]/v_lim[i_jammed]
data_2 = v_ego[i_synch]/v_lim[i_synch]
data_3 = v_ego[i_free]/v_lim[i_free]
data = [data_1, data_2, data_3]

fig, ax = plt.subplots(1,1)
ax.boxplot(data)
ax.set_xticklabels(['jammed','synchronized','free'])
ax.set_ylabel('v/vlim')
```

- The code will not run as there is an error in the data variable creation
- The code will create a boxplot of the data
- …

9

Given the following features and code in order :





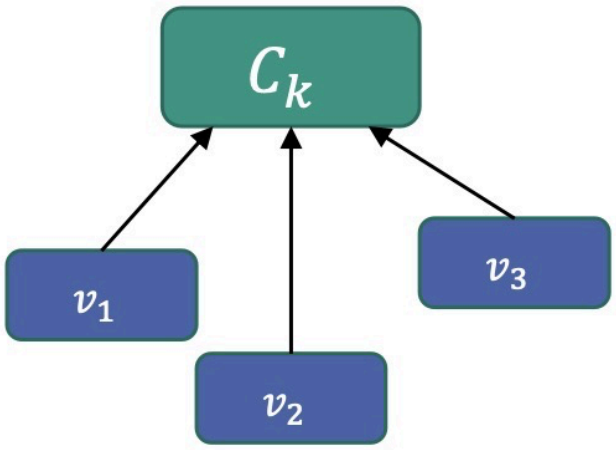

```
o, v and v_l are known
i_j = [ ]
i_synch = [ ]
i_ff = [ ]
for (i in o):
        if( v[i]/v_l[i] <= x_1 and o_1[i] == x_2):
                i_j.append(i)
        elif( v[i]/v_l[i] > y_1 and o_1[i] == y_2):
                i_synch.append(i)
        elif( v[i]/v_l[i] >= z_1 and o_1[i] == z_2):
                i_ff.append(i)
```

Which rules should be chosen to

- $x_1, y_1 = 0,5 \quad z_1 = 1$
- $x_1, y_1, z_1 = 0,5$
- $x_2, y_2 = 1 \quad z_2 = 0$
- $x_2, y_2 = 0 \quad z_2 = 1$

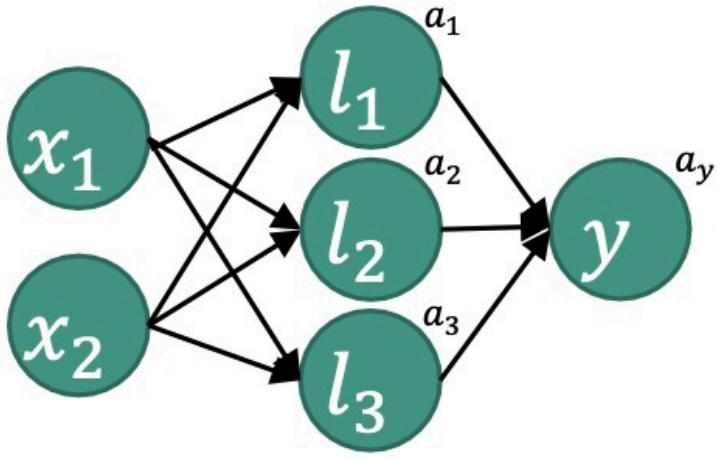
10

The underlying model has feature inter-dependencies. What model is shown here?



- This shows a Bayesian Network
- This shows a Naive Bayesian Classifier
- This shows a Neural Network

11

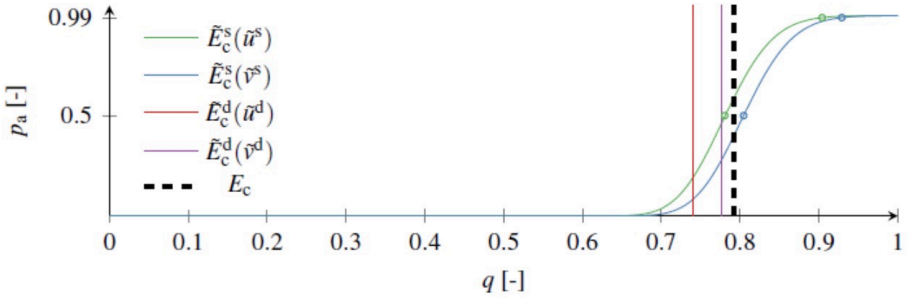


According to the graphic, which of the statements are true?

- This model has 3 neurons
- This model has 1 hidden layer
- This model is a Deep Neural Network
- This model is a Neural Network
- The activation functions a need to be linear

12

Which of the following are true in regards to the graphic?



- The driver needs a SoC of 0,8 to reach his destination
- Deterministic models overestimate the energy consumption
- Deterministic models underestimate the energy consumption
- According to the stochastic model, to reach an attainability value of p = 0,99 the SoC needs to be at 0,9
- According to the stochastic model, at a SoC of 0,8 50% attainability is reached

13?

Monte Carlo

- Monte Carlo is distribution independent
- Unscented Transformation uses randomly distributed sigma points
- Monte Carlo is highly dependent on the sampling procedure
- …

14?

Convolution

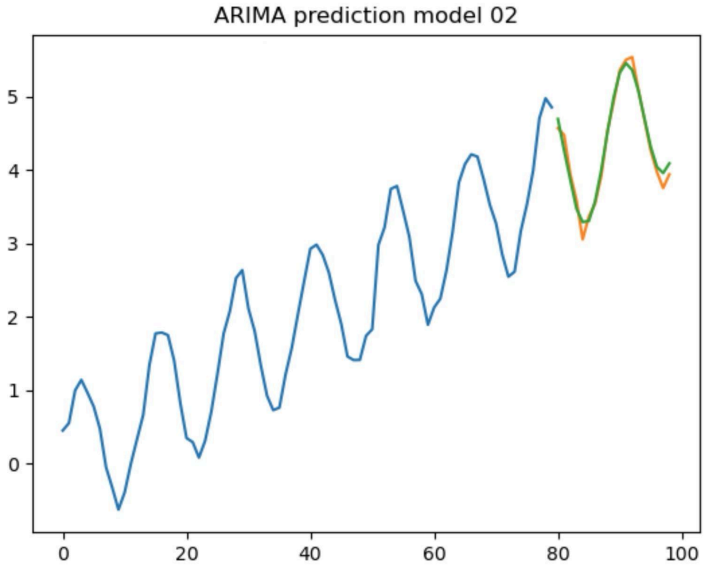
- Convolution needs low-dependency features
- Convolution with many features will always resemble a Gaussian normal distribution because of the central limit theorem
- To use added feature probabilities one needs to compute the Cumulative Density Functions (CDFs) from the PDFs
- …

15?

What are true about Online learning/Federated learning

- Sparse learning limits the memory used to train the model
- Federated learning distinguishes between important and unimportant information. The model is not retrained for every timestep.
- …

16?



(+ code given)

What is true about ARIMA and this graphic?

- the model cannot predict the data well as it does not learn the the relevant data pattern
- The model is able to predict the data well
- …

17?

Which of the following statements is true about CRSP?

- allows for stochastic forecasts to be compared to deterministic ones
- …

18

Knowing the traffic phase and whether a vehicle is in front of the ego vehicle, one can calculate the likelihood for being the synchronized traffic phase in Bayesian Statistics by

- dividing the total observations of synchronized traffic by all observations
- dividing the synchronized and car in front observations by all synchronized traffic phase observations
- dividing the synchronized and car in front observations by all car in front observations
- …

19

With 2 degrees of freedom, what is largest value α for which you can refuse the H0 hypothesis?

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

- $\alpha = 0.9$
- $\alpha = 0.95$
- $\alpha = 0.1$
- $\alpha = 0.01$

20

Which test can you use to check whether two observed variables are independent from each other

- Kruskal-Wallis Test
- Chi-Squared Test
- …
- …

21

Which metrics are robust?

- Mean Absolute Error
- Mean Percentage Error
- Median Percentage Error
- None of the above

22

The advantages of Recursive Linear Regression model

- can deal with infinite time-series
- cannot take more than 2 features of input as it is linear
- the memory limit is known in advance
- is not influenced by poor excitation

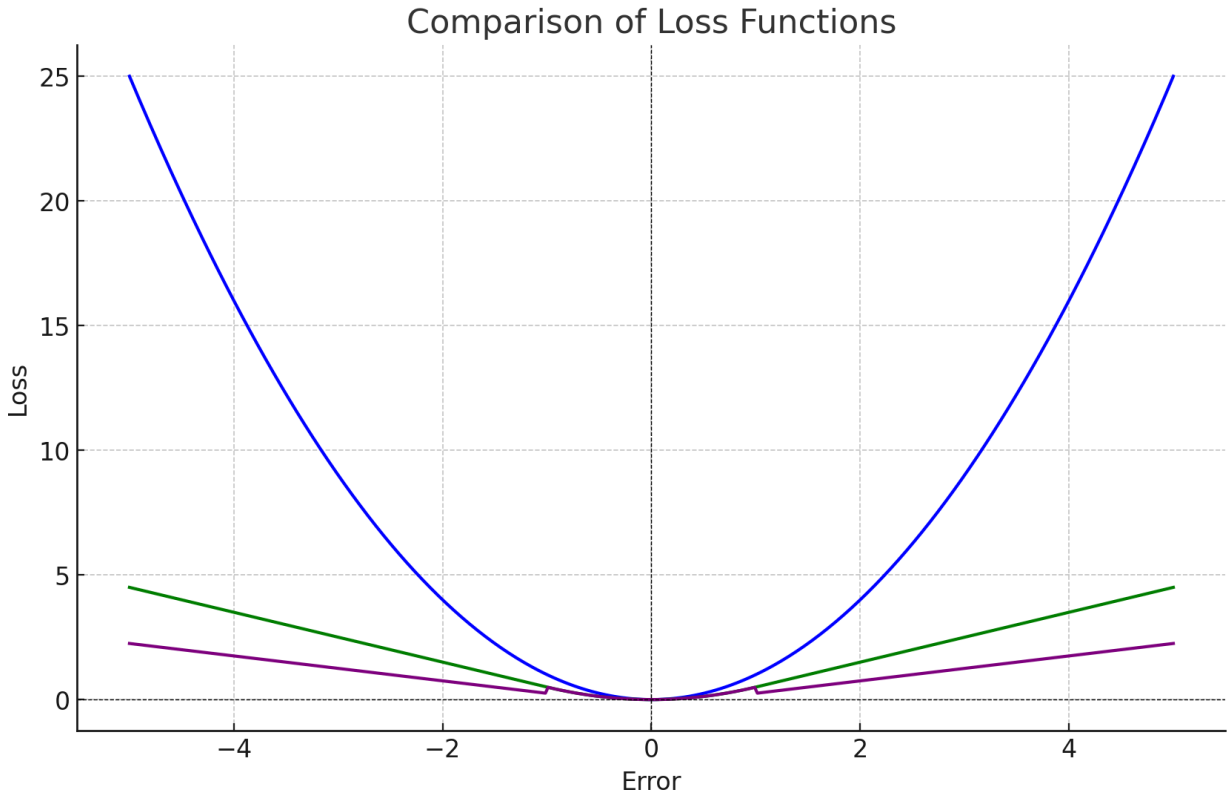
23

What is labeled data?

- Data coming from an API which offers a description of the data
- Data for which the input and output are known
- …

24

Which loss function is the most robust?



- Green loss function
- Blue loss function
- Purple loss function
- It is not possible to tell from the graphic

25

What does the following code do?
```
*consider all variables not initialized in the following code known*
    (1) mlr = LinearRegression()
    (2) mlr.fit(X1,y1)
    (3) mlr_train_y = mlr.predict(X1)
    (4) mlr_train_res = mlr_train_y-y1

    (5) svr = SVR()
    (6) svr.fit(X1, y1, kernel="rbf", C=reg, epsilon=eps)
    (7) svr_train_y = svr.predict(X1)
    (8) svr_train_res = svr_train_y-y1
```

- Both models predict a variable y and their residuals are calculated
- The model in line 1 is not initialized correctly
- The model in line 5 is not initialized correctly
- …

26

What does the following code do?
```
mlr_1 = LinearRegression()
mlr_1.fit(X1,y1)

mlr_2 = LinearRegression()
mlr_2.fit(X2, y2)

is_res_1 = y_1 - mlr_1.predict(X1)
is_res_2 = y_2 - mlr_2.predict(X2)

oos_res_1 = y_2 - mlr1_predict(X1)
oos_res_2 = y_1 - mlr_2.predict(X2)

```

- The model's out of sample residuals are calculated and stored in variables
- The model's in sample residuals are calculated
- None of the above
- …

27

What is true about the Dynamic Linear Regression?

- Uses an ARIMA model as a loss function
- Changes its parameters after a certain, therefore being “dynamic”
- Uses normal and time-series data
- …

28

What is true about Decision Trees?

- Decision Trees can only be used for classification
- Multiple Decision Trees of different lengths make up a Random Forest
- Decision Trees tend to overfit data
- Random Forests tend to overfit data

29

Which of the following is true about Support Vector Machines (SVMs)?

- SVMs can classify non-linear data using the kernel trick
- SVMs cannot be used for estimation through regression unless a big enough ϵ tube is specified
- …

Solutions

*excs -> code exercise as of Summer 2024

Question Number	Answer
1	JSON
2	wrong plot dimensions => error in line 5
3	PDF: relative continuous, PMF: discrete
4	red: higher bandwidth (see excs* 4_01)
5	gray box
6	Does not minimize the sum of squared errors
7	plots elbow plot (see excs. 12_01)
8	creates boxplot (see excs 7_01)
9	2,3
10	Naive Bayesian Classifier
11	2,4,5
12	1,2,4,5
13	
14	
15	
16	
17	
18	sync and car in front / all sync
19	$\alpha = 0.1$
20	Kruskal; Chi-squared
21	Median Percentage Error
22	infinite time-series; known memory limit
23	known input and output
24	Purple function

25	3
26	2
27	
28	2,3
29	1