Exam

When and Where:

- **When:** March, 26. 08:00 - 10:00 (duration is 90 min, starting from 08:30.) **Be there at 08:00!**
- **Where:** Zelt auf dem Forum (großes Zelt)

Exam

Karlsruhe Institute of Technology

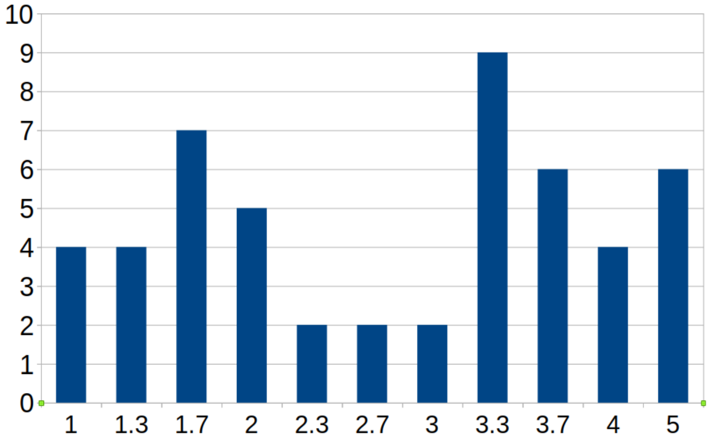
When and Where:

- **When:** March, 26. 08:00 - 10:00 (duration is 90 min, starting from 08:30.) **Be there at 08:00!**
- **Where:** Zelt auf dem Forum (großes Zelt)

Rules:

- There will be 30 questions - see examples on following slides.
- You can bring a cheat-sheet. One A4-sheet, handwritten, you can write on both sides!
- There will **not** be python in the exam.

Overview Results Last Year



This was without bonus!

Example Question 1

Question:
Under which assumptions is the least squares objective from linear regression equivalent to a maximum likelihood objective?

Example Question 1

Question:
Under which assumptions is the least squares objective from linear regression equivalent to a maximum likelihood objective?

Answer:
Gaussian likelihood with linear mean and constant noise.

Example Question 2

Question:
Write down the objective of linear binary logistic regression. The samples are given by $\boldsymbol{x}_i$ and the labels by $c_i \in \{0, 1\}$.
How is $p(c_i|\boldsymbol{x}_i)$ assumed to be distributed in binary logistic regression?

Example Question 2

Question:
Write down the objective of linear binary logistic regression. The samples are given by $\boldsymbol{x}_i$ and the labels by $c_i \in \{0, 1\}$.
How is $p(c_i|\boldsymbol{x}_i)$ assumed to be distributed in binary logistic regression?

Answer:

$$\text{argmax}_{\boldsymbol{w}} \sum_{i=1}^{N} c_i \log(\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)) + (1 - c_i)\log(1 - \sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$$

Logistic regression assumes $p(c_i|\boldsymbol{x}_i)$ to be Bernoulli distributed

Example Question 3

Karlsruhe Institute of Technology

Question:

What are the hyperparameters for choosing the model complexity for each of the following algorithms. Name at least one hyperparameter for every algorithm.

- Neural Networks
- Support Vector Machines
- Gaussian Processes
- Decision Trees?

Example Question 3

Question:

What are the hyperparameters for choosing the model complexity for each of the following algorithms. Name at least one hyperparameter for every algorithm.

- Neural Networks
- Support Vector Machines
- Gaussian Processes
- Decision Trees?

Answer:

- Neural Networks: number of layers, number of neurons
- Support Vector Machines: which features to choose (also kernel bandwidth included), regularization
- Gaussian Processes: kernel bandwidth , prior
- Decision Trees: maximum depth, number of leaves

Example Question 4

Question:

Name at least two advantages and two disadvantages of decision trees.

Example Question 4

Question:

Name at least two advantages and two disadvantages of decision trees.

Answer:

Advantages:

- Applicable to both regression and classification problems.
- Handle categorical predictors naturally.
- Computationally simple and quick to fit, even for large problems.
- No formal distributional assumptions (non-parametric).
- Can handle highly non-linear interactions and classification boundaries.
- Very easy to interpret if the tree is small.

Disadvantages:

- Accuracy: current methods, such as support vector machines and ensemble classifiers often have 30% lower error rates than CART.
- Instability: If we change the data a little, the tree picture can change a lot. So the interpretation is not as straightforward as it appears.

Example Question 5

Question:
Which data-structure is usually used to efficiently implement k-Nearest Neighbors? Name the main steps in building that data-structure.

Example Question 5

Question:
Which data-structure is usually used to efficiently implement k-Nearest Neighbors? Name the main steps in building that data-structure.

Answer:
KD-Trees
Building the tree:

- Choose dimension (e.g., longest hyper-rectangle).
- Choose median as pivot
- Split node according to (pivot, dimension)

Example Question 6

Question:

First, explain the intuition behind slack-variables in support vector machine training.

Second, for a single data-point $(\boldsymbol{x}_i, c_i)$ the margin condition with slack variable ξ_i is given as

$$c_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i.$$

- Assuming $0 \leq \xi_i \leq 1$, is $\boldsymbol{x}_i$ classified correctly?
- Assuming $\xi_i > 1$, is $\boldsymbol{x}_i$ classified correctly?

Example Question 6

Question:
First, explain the intuition behind slack-variables in support vector machine training.
Second, for a single data-point $(\boldsymbol{x}_i, c_i)$ the margin condition with slack variable ξ_i is given as

$$c_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i.$$

- Assuming $0 \leq \xi_i \leq 1$, is $\boldsymbol{x}_i$ classified correctly?
- Assuming $\xi_i > 1$, is $\boldsymbol{x}_i$ classified correctly?

Answer: One of the following

- Allow violation of margin condition
- Act as regularization, avoid over-fitting
- Make constraint optimization problem solvable, even if the data is not linearly separable

Other Questions

- Yes, (but margin is violated)
- No

Example Question 7

Question:

You are given the following optimization problem:

$$\underset{a}{\operatorname{argmax}}\ a^2 h$$
$$s.t.\ \ S_{max} \geq 2a^2 + 4ah$$

Write down the Lagrangian. Derive the optimal value for a depending on your lagrangian multiplier.

Example Question 7

Karlsruhe Institute of Technology

Question:

You are given the following optimization problem:

$$\underset{a}{\operatorname{argmax}}\ a^2 h$$
$$s.t.\ \ S_{max} \geq 2a^2 + 4ah$$

Write down the Lagrangian. Derive the optimal value for a depending on your lagrangian multiplier.

Answer:

$$L = a^2 h + \lambda(S_{max} - 2a^2 - 4ah)$$
$$\frac{dL}{da} = 2ah - 4\lambda a - 4\lambda h$$
$$a^* = \frac{4\lambda h}{2h - 4\lambda}$$

(The sign in the Lagrangian may change. But the following equations should be correct according to the Lagrangian.)

Example Question 8

Question:
What is the key idea behind second order optimization methods? What are their benefits? Why are second order optimization methods usually not applicable for Deep Neural Networks?

Example Question 8

Question:
What is the key idea behind second order optimization methods? What are their benefits? Why are second order optimization methods usually not applicable for Deep Neural Networks?

Answer:

- To use the second derivative (Hessian) of the objective and directly step towards the minimum of the quadratic approximation.
- No learning rate needs to be tuned and they need fewer function evaluations.
- Because the Hessian is huge and needs to be inverted.

Example Question 9

Question:
Why is it not feasible to use fully connected layer for images? How do convolutional neural networks solve this problem and which property of an image do they exploit.

Example Question 9

Question:
Why is it not feasible to use fully connected layer for images? How do convolutional neural networks solve this problem and which property of an image do they exploit.

Answer:

- They have to many parameters
- less parameters by parameter sharing
- They exploit spatial structure

Example Question 10

Question:

Gaussian Processes(GP) are also referred to as a ”Bayesian Kernel Regression” approach. Why?

Example Question 10

Question:
Gaussian Processes(GP) are also referred to as a ”Bayesian Kernel Regression” approach. Why?

Answer:
You can derive Gaussian Processes by using the Kernel Trick from Bayesian Linear Regression, as in Kernel Regression.