Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning
2016

Prof. J. Buhmann

Final Exam

January 24th, 2017

First and Last name: ______________________________

Student ID (Legi) Nr: ______________________________

Signature: ______________________________

General Remarks

- Please check that you have all ?? pages of this exam.
- There are 180 points, and the exam is 180 minutes. **Don't spend too much time on a single question!** The maximum number of points is not required for the best grade!
- Remove all material from your desk which is not permitted by the examination regulations.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your **student-ID**-number on top of each supplementary sheet.
- Immediately inform an assistant in case you are not able to take the exam under regular conditions. Later complaints are not accepted.
- Attempts to cheat/defraud lead to immediate exclusion from the exam and can have judicial consequences.
- Please use a black or blue pen to answer the questions.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Regression, Kernels, GPs	60		
2	Classification, Ensembles & Mixtures	60		
3	HMMs, Neural Nets & Evaluation	60		
Total		180		

Grade: ....

1 Regression, Kernels, GPs (60 pts)

1.1 Regression - Warm up

We have the following linear regression model

$$Y = 3 - 2X + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $X \sim \mathcal{N}(1, 4)$ and ϵ and X are independent. Note that $T \sim \mathcal{N}(a, b)$ signifies that the random variable T follows a normal distribution with mean a and variance b.

a) Calculate the mean and variance of Y

5 pts □

....

....

1.2 Lasso-Ridge Regression

Which of the following claims are true/false? *(2 points per correct answer, -2 points per incorrect answer, 0 per blank, non-negative total points in any case)*

10 pts □

a) The lasso estimator is biased.
[ ] True [ ] False

b) The lasso cost function is convex.
[ ] True [ ] False

c) Ridge regression is better than Lasso since it gives us unbiased estimators.
[ ] True [ ] False

d) Lasso always has a closed-form solution.
[ ] True [ ] False

e) In Ridge regression, as the regularization parameter increases, the regression coefficients decrease.
[ ] True [ ] False

1.3 High-Dimensional Regression

We have the following regression model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, $\epsilon \in \mathbb{R}^n$ and $\mathrm{Var}(\epsilon) = \sigma^2 \mathbf{I}$

a) Calculate $\hat{\beta}$ using least squares. (Please do the derivation)

4 pts □

....

....

....

b)Which assumption on $\mathbf{X}$ do you need to invert $\mathbf{X}^T\mathbf{X}$?

2 pts □

....

....

c) Calculate the expected value and variance of $\hat{\beta}$.

4 pts □

....

....

....

1.4 Kernels

Unless stated otherwise, assume $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, with $\mathbf{x} = (x_1, \ldots, x_d)$ and the same for $\mathbf{y}$. The symbol $\|\cdot\|$ denotes the Euclidean norm and $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ denotes the inner product in the respective space.

a) Assume that $k_0(\mathbf{x}, \mathbf{y})$ *is not* a kernel and $k_1(\mathbf{x}, \mathbf{y})$ *is* a kernel. Can you say anything about whether $k_2(\mathbf{x}, \mathbf{y}) := k_0(\mathbf{x}, \mathbf{y}) - k_1(\mathbf{x}, \mathbf{y})$ is a kernel or not? Or does it depend on specific forms of k_0 and k_1? Motivate your answer.

7 pts □

....

....

....

b) Imagine you are asked to prove that a real-valued function $k(\mathbf{x}, \mathbf{y})$ is a kernel. What is wrong in the following "proof"?

"First, check that $k(\mathbf{x}, \mathbf{y})$ is symmetric. Then let us show that for *any* pair of vectors $\mathbf{x}, \mathbf{y}$ the Gram matrix

$$G := \begin{pmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{y}) \\ k(\mathbf{y}, \mathbf{x}) & k(\mathbf{y}, \mathbf{y}) \end{pmatrix}$$

is positive semi-definite. Then k is a kernel."

4 pts □

....

....

1.5 SVMs

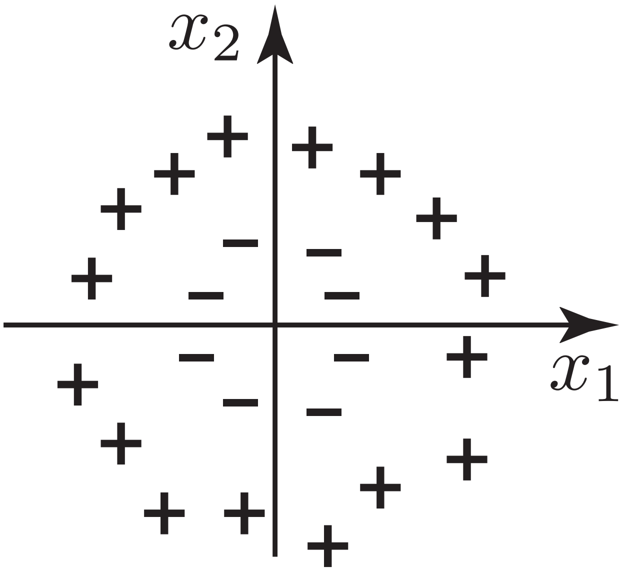
a) Name two ways in which the formulation of basic linear hard-margin SVM can be changed in order to deal with (1) structural non-separability of the training set and (2) possibly noise-induced non-separability of the training set.

3 pts □

....

....

b) Imagine you are given a (linearly non-separable) dataset:



Which feature map of $\mathbf{x} = (x_1, x_2)$ would you choose to classify this dataset with zero-training error with a linear classifier?

3 pts □

....

....

c) For the feature map of the previous task, write down the respective kernel (assume $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ in terms of x_i and y_i:

5 pts □

....

....

1.6 Gaussian Processes

Consider a Gaussian process:

$$\mathbf{y} = \mathbf{f}(\mathbf{X}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{d \times n}$ and $\mathbf{f}$ is drawn from a Gaussian processes:

$$\mathbf{f}(\mathbf{X}) \sim \mathcal{GP}(m(\mathbf{X}), k(\mathbf{X}, \mathbf{X}')),$$

where $m(\mathbf{X})$ is the mean function and $k(\mathbf{X}, \mathbf{X}')$ is the kernel.

a) Derive the distribution of $\mathbf{y}$:

5 pts □

....

....

....

....

....

b) Derive the joint distribution of all the data $\mathbf{y}$ and a new observation y_{n+1}:

4 pts □

....

....

....

....

....

c) Suppose $\boldsymbol{\Theta}$ are the hyperparameters that parameterize the kernel function $k(\mathbf{X}, \mathbf{X})$. Describe a method to optimize such hyperparamters.

4 pts □

....

....

....

....

2 Classification, Ensembles & Mixtures (60 pts)

2.1 Classification

a) Show how the posterior probability of class membership given data can be rewritten as a logistic function. Don't assume a specific probability distribution.

5 pts □

....

....

....

....

b) What is the argument of the function? What does it mean when its sign changes? How can it be used for classification?

5 pts □

....

....

....

....

c) Given a cost function $J(\theta)$ (θ are the learnable parameters, or weights, of a classifier), outline the gradient descent algorithm with a fixed learning rate η.

5 pts □

....

....

....

....

d) Outline a 2^{nd} order algorithm to find the optimal learning rate.

5 pts □

....

....

....

....

2.2 Bagging and Boosting - Warm up

a) Explain how bagging reduces the covariance of the outputs of weak learners.

3 pts □

....

....

. . . .

b) State two essential differences between bagging and boosting.

3 pts □

. . . .

. . . .

c) State two factors that may influence the performance of AdaBoost.

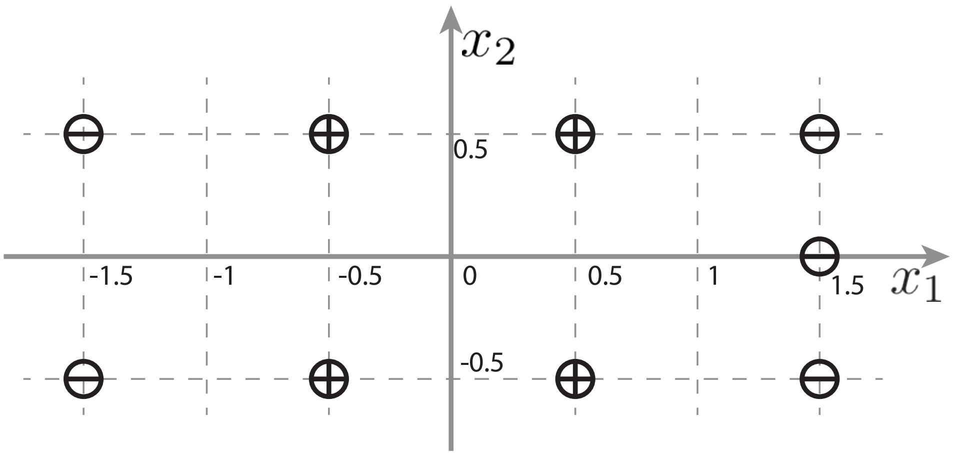
3 pts □

. . . .

. . . .

2.3 AdaBoost

In this example, we are going to apply the AdaBoost algorithm to classify nine data points in two dimensional space by hand, as shown by the following figure.



Each "$\oplus$" represents a data point with label $y = 1$, each "$\ominus$" represents a data point with label $y = -1$. In detail, the nine data points are listed in the following table,

data point index	1	2	3	4	5	6	7	8	9
y	-1	-1	-1	-1	-1	1	1	1	1
x_1	-1.5	-1.5	1.5	1.5	1.5	-0.5	-0.5	0.5	0.5
x_2	0.5	-0.5	0.5	0	-0.5	0.5	-0.5	0.5	-0.5

For the weak classifiers, we use decision stump in the following form,

$$\text{stump}(\mathbf{x}|k, \theta, \gamma) = \begin{cases} +\gamma, & \text{if } x_k > \theta, \\ -\gamma, & \text{otherwise} \end{cases}$$

where $k \in \{1, 2\}$ indicates which dimension to use, θ is an *integer* indicating the threshold, $\gamma \in \{-1, 1\}$ representating the labelling "direction".

a) Write down the first two weak classifiers chosen by AdaBoost.

3 pts □

....

....

....

....

b) Are the first two weak classifiers sufficient to correctly classify the nine data points? If so, give one possible example of the final classifier ($\hat{c}_B(\mathbf{x})$); If not, at least how many weak classifiers are needed to correctly classify the nine data points? Write down one possible form of the extra weak classifier(s).

4 pts □

....

....

....

....

2.4 Mixture Models - Warm up

Which of the following claims are true/false? *(2 points per correct answer, -2 points per incorrect answer, 0 per blank, non-negative total points in any case)*

8 pts □

a) The k-means algorithm is computationally more efficient than clustering with Gaussian mixture models.
[ ] True [ ] False

b) Direct maximization of log-likelihoods is intractable for mixture models.
[ ] True [ ] False

c) The Expectation Maximization (EM) method increases log-likelihood in each iteration.
[ ] True [ ] False

d) EM converges to the global maximum of log-likelihood.
[ ] True [ ] False

2.5 k-means algorithm and initialization

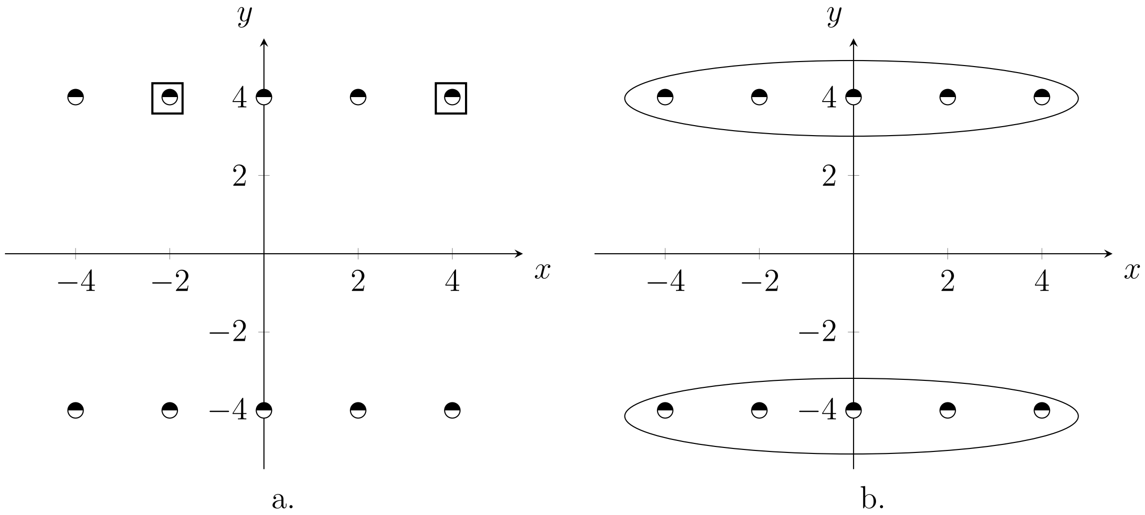


Figure 1: Data set for clustering; In figure a., points inside rectangles are initial centres. In figure b., we aim at retrieving determined clusters.

a) Assuming that initial centres are points $(x = -2, y = 4)$ and $(x = 4, y = 4)$, draw final clusters that are obtained by k-means clustering $(k = 2)$ on Figure ??.a.

3 pts □

b) Suggest an initialization on Figure ??.b. such that k-means converges to determined clusters in Figure ??.b.

3 pts □

2.6 Bayesian Learning

Consider a polynomial of degree 14 used to fit a sample y_i drawn from an unknown probability distribution:

$$\begin{aligned}\hat{y}_i(x) &= \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \ldots + \theta_{14} x_i^{14} \\ &= \boldsymbol{\phi}_i \boldsymbol{\theta}\end{aligned}$$

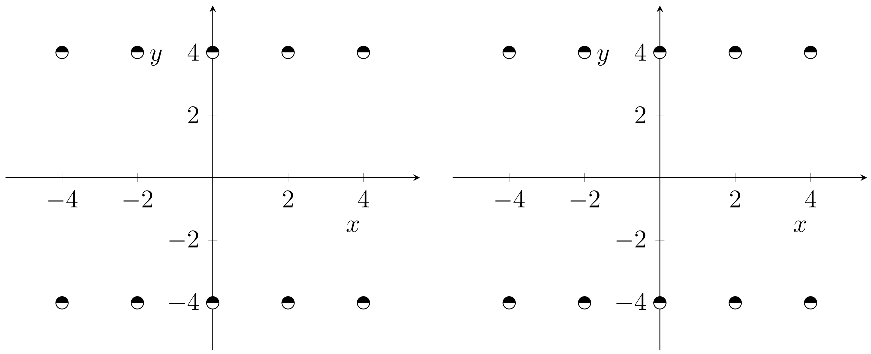


Figure 2: Practice figure for clustering

where $\hat{y}_i(x_i)$ is the estimated response, x_i is the independent variable and $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots \theta_{14})^T$ are the polynomial parameters. $\boldsymbol{\phi}_i$ is given by $\boldsymbol{\phi}_i = (1, x_i, x_i^2, \ldots, x_i^{14})$. The likelihood function for n observations is given by:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \exp\Big(-\left(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\right)\Big),$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)^T$.

a) Derive the closed-form solution of the maximum likelihood estimator of the polynomial parameters $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$:

5 pts □

. . . .

. . . .

. . . .

. . . .

b) Now consider adding a regularizer:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) \exp\left(-\delta \boldsymbol{\theta}^T \boldsymbol{\theta}\right),$$

where δ is a scalar.

Derive the maximum a posteriori estimate $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$:

5 pts □

. . . .

. . . .

. . . .

. . . .

3 HMMs, Neural Nets & Evaluation (60 pts)

3.1 Hidden Markov Models

a) Given an HMM characterised as below, evaluate the probability of an observation sequence $O = v_1, v_2, v_1$. Note that each row sums up to 1 in tables for A and B, indicating $P(S_2|S_1) = 0.2$. Use the following quantity $\alpha_t(i) = P(O_1\,O_2\,\cdots O_t,\ q_t = S_i|\lambda)$ to compute the probability. You can assume that there are only 2 states and 2 observations.

S_1	S_2
0.4	0.6

π
initial prob.

	S_1	S_2
S_1	0.8	0.2
S_2	0.7	0.3

A
state transition

	v_1	v_2
S_1	0.9	0.1
S_2	0.25	0.75

B
observation

10 pts □

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

b) Define forward and backward variables as $\alpha_t(i) = P(O_1\,O_2\,\cdots O_t,\ q_t = S_i|\lambda)$ and $\beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T|q_t = S_i, \lambda)$, where λ is the HMM. Given $\gamma_t(i) = P(q_t = S_i|O, \lambda)$, prove that $\gamma_t(i) = \alpha_t(i)\beta_t(i)/P(O|\lambda)$. Explicitly state any assumptions you use. Also, use the equality $O = O_1\,O_2\,\cdots\,O_T$.

8 pts □

. . . .

. . . .

. . . .

. . . .

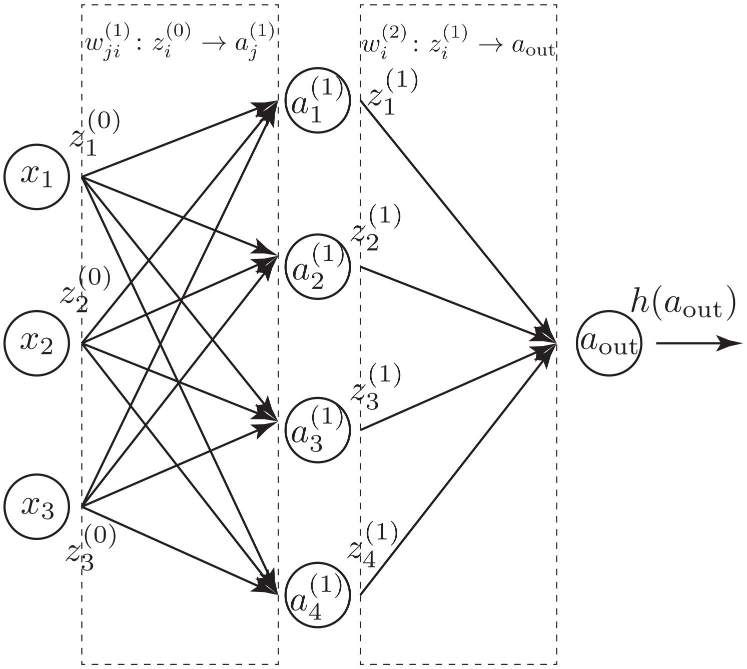
. . . .

. . . .

3.2 Neural Networks

Unless stated otherwise, we use the notation of the tutorial. Numbers are given so that you do not need a calculator. "Explain briefly" means no rigorous proof is needed, but rather a clear hint of the core idea you used.

Consider the following neural network:



It consists of the input (i.e. 0^{th}) layer $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ and the hidden (i.e. 1^{st}) layer. These layers have linear activation functions $z_i^{(j)}(\cdot)$. Note there are no additive constants in activations.

Output activation function is sigmoid: $h(a_{\text{out}}) = \frac{1}{1+\exp(-a_{\text{out}})}$.

a) Which task is this neural net most likely designed for? Explain briefly.

5 pts □

. . . .

. . . .

....

b) Given that for input $\mathbf{x} = (1, -1, -0.5)$ the output is $h(a_{\text{out}}) = 1/2$, what would be the output for $\mathbf{x}' = (-2, 2, 1)$? Explain briefly.

5 pts □

....

....

....

c) For the above neural network, assuming the loss function to be denoted L. To compute which gradient is the final goal of backpropagation algorithm?

4 pts □

....

....

d) Which issue is addressed by the *simulated annealing procedure*?

1 pts □

....

e) It is known that *a perceptron with one hidden layer can approximate infinitely well every continuous function*. Why is this not typically done in practice?

5 pts □

....

....

....

3.3 Evaluation

Which of the following claims are true/false? *(2 points per correct answer, -2 points per incorrect answer, 0 per blank, non-negative total points in any case)*

8 pts □

a) Leave-one-out cross-validation provides a biased estimation of the true prediction error.
[ ] True [ ] False

b) The variance of the estimated error is high for leave-one-out cross-validation.
[ ] True [ ] False

c) The Bootstrap method sacrifices samples for the sake of validation.
[ ] True [ ] False

d) The Bootstrap estimation of classification errors is too pessimistic due to sampling with replacement.
[ ] True [ ] False

3.4 K-fold Cross Validation

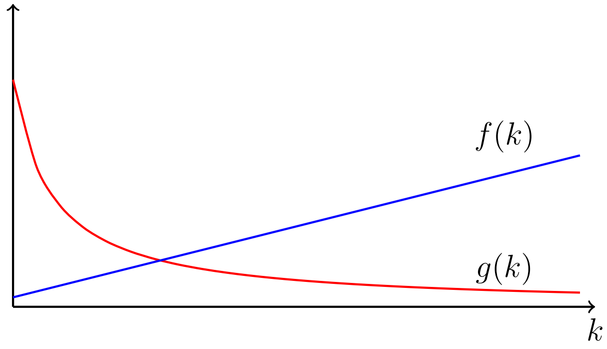


Figure 3: The bias-variance tradeoff for k-fold cross validation

Figure ?? illustrates the bias-variance tradeoff of the estimated prediction error of k-fold cross validation for different choices of k. There are two trends in the plot: the decreasing trend $g(k)$, and the increasing trend $f(k)$. Mark the right choice that completes following sentences. *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

a) The trend $g(k)$ shows (bias [ ], variance [ ]). It decreases for large k because (less samples are sacrificed for validation [ ], training sets for different folds are more correlated [ ])

4 pts □

b) The trend $f(k)$ shows (bias [ ], variance [ ]). It increases for large k because (less samples are sacrificed for validation [ ], training sets for different folds are more correlated [ ])

4 pts □

3.5 Model Selection

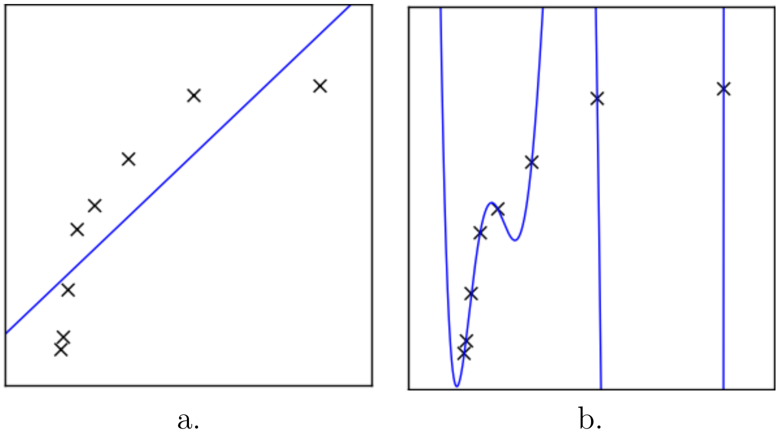


Figure 4: Model Selection

We used two different approaches of assessment for model selection: 5-fold cross validation, and bootstrapping. Figure ?? represents selected models by these approaches. As you see, the model a. is relatively simpler than model b. Which model is selected by cross validation? Why?

6 pts □

. . . .

. . . .

. . . .

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet