January 31, 2024
Prof. Dr. Marie Düker, Jorge Weston

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Exam Selected Topics in Mathematics of Learning

Full Name: ....

Matr. Number: ....

Study Program: ....

Signature: ....

General Remarks

- You have **60 minutes**. The exam consists of this page (P1) with instructions, pages (P2-P8) with the exam questions. Check for completeness.
- Write with a **black** or **blue** pen. Don't use a pencil, ink, or anything else that can be easily erased or becomes blurry. If you want to correct something, **cross it out**, such that it is **still clearly readable** what has been written.
- If you are feeling sick, we recommend to leave the exam now. If you are feeling sick after you have started the exam, raise your hand and let us know. You will have to visit the university's medical officer for a certificate.
- **No** additional tools are allowed! You are **not** allowed to use a cheat sheet!
- Especially, it is **not allowed to use any electronic device** without the examiner's explicit permission. This includes notebooks, smart phones, tablets, or any "smart" device like watches. Turn off all your electronic devices.
- It is not allowed to communicate with anyone except the examiners during the exam. It is not allowed to look at somebody else's exam sheet.
- If one of the above rules is violated, the examiners might grade your exam with **5,0 (failed)** immediately, and might set further disciplinary measures in motion.
- **We expect that each answer is substantiated properly, except explicitly stated otherwise.**

Exercise	**1.**	**2.**	**3.**	**4.**	**Total**
Points	**12**	**12**	**12**	**12**	**48**
Achieved					

Exercise 1. **(/12 P.)**

Question 1: In LASSO regression, if the regularization parameter $\lambda = 0$, then which of the following is true? (/2 P.)

A. This LASSO model can be used for feature selection.

B. The loss function is the same as the ridge regression loss function.

C. The loss function is the same as the ordinary least square loss function.

D. Large coefficients are penalized.

Question 2: Ridge regression is... (/2 P.)

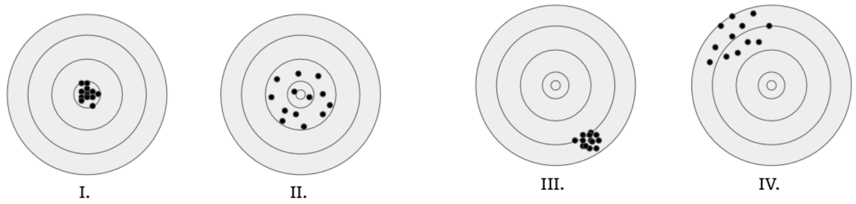
A. ...a way to perform feature selection, as ridge regression encourages weights to be exactly zero.

B. ...has the same solution as the ordinary least square loss function.

C. ...produces sparser results (more zero weights) than LASSO regression.

D. ...a method in which bias tends to increase, and variance tends to decrease, as we increase the regularization parameter λ.

Question 3: The following graphic will be used as a representation of bias and variance. Imagine that a true/correct model is one that always predicts a location at the center of each target (being farther away from the center of the target indicates that a model's predictions are worse). We retrain a model multiple times, and make a prediction with each trained model. For each of the targets, determine whether the bias and variance is low or high with respect to the true model. (/2 P.)



Subplot I: How are bias and variance related to the true model?

A. High bias, High variance

B. High bias, Low variance

C. Low bias, High variance

D. Low bias, Low variance

Subplot II: How are bias and variance related to the true model?

A. High bias, High variance

B. High bias, Low variance

C. Low bias, High variance

D. Low bias, Low variance

Subplot III: How are bias and variance related to the true model?

A. High bias, High variance

B. High bias, Low variance

C. Low bias, High variance

D. Low bias, Low variance

Subplot IV: How are bias and variance related to the true model?

A. High bias, High variance

B. High bias, Low variance

C. Low bias, High variance

D. Low bias, Low variance

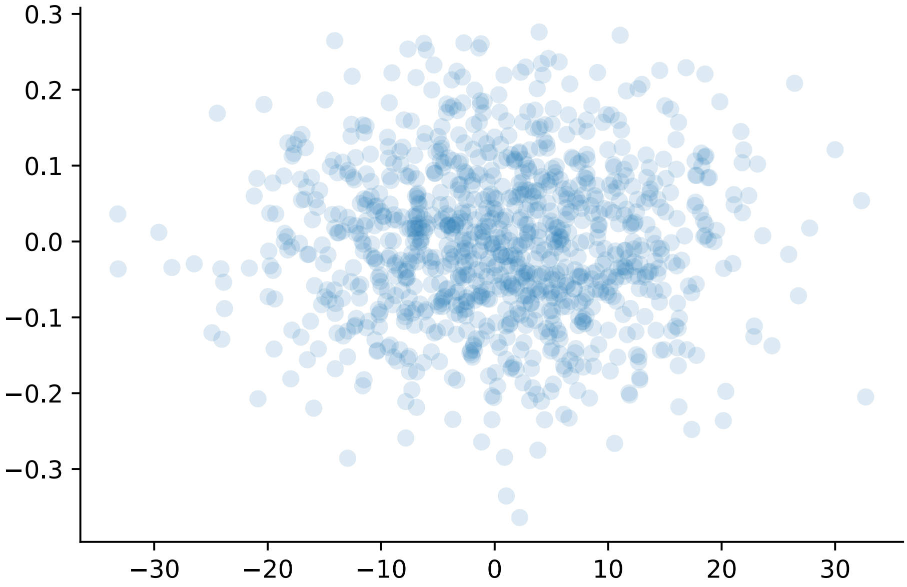
Question 4: Which of the following functions is <u>not</u> a norm: (/2 P.)

A. $\|\beta\|_0 = \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}}$

B. $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

C. $\|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$

Question 5: Below are 1000 sample points drawn from a two-dimensional multivariate normal distribution. Which of the following matrices could (without extreme improbability) be the covariance matrix of the distribution? (Pay attention to the numbers on the axes!) (/2 P.)



A.

$$\Sigma = \begin{pmatrix} 100 & 0 \\ 0 & 0.01 \end{pmatrix}$$

B.

$$\Sigma = \begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}$$

C.

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

D.

$$\Sigma = \begin{pmatrix} -10 & 0 \\ 0 & -0.1 \end{pmatrix}$$

Question 6: For $\lambda > 0$, what estimation technique is the following function used for (/2 P.)

$$u \mapsto \operatorname{sign}(u)(|u| - \lambda)\mathbb{1}_{\{|u| \geq \lambda\}}.$$

A. Hard-thresholding

B. Soft-thrsholding

Solution

Question 1 C is correct.
Question 2 D is correct.
Question 3 I low low, II low high, III high low, IV high high.
Question 4 A is correct.
Question 5 A is correct.
Question 6 B is correct.

Exercise 2. (/12 P.)

(a) Let X be a random variable with $\mathrm{E}\,X = 0$. Suppose that the moment-generating function of X^2 is bounded at some point, that is,

$$\mathrm{E}\exp(X^2) \leq 2.$$

Show that X satisfies the two-sided tail bound

$$\mathrm{P}(|X| > t) \leq 2\exp(-t^2) \text{ for all } t \geq 0.$$

(/4 P.)

(b) Prove that if X is a non-negative random variable with expectation $\mathrm{E}[X]$, then for all $t > 0$, we have $\mathrm{P}[X \geq t] \leq \mathrm{E}[X]/t$. (/4 P.)

(c) Recall Chernoff's inequality: Let X_i be independent Bernoulli random variables with success probability p_i. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathrm{E}[S_N]$. Then, for any $t > \mu$, we have

$$\mathrm{P}(S_N \geq t) \leq \exp(-\mu)\left(\frac{\exp(1)\mu}{t}\right)^t.$$

Consider 200 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 150. Use Chernoff's inequality. (/4 P.)

Solution

(a) It holds that

$$P(|X| > t) = P(e^{X^2} > e^{t^2}) \leq \frac{2}{e^{-t^2}}.$$

(we just apply Markov in the last step)

(b) It holds that

$$P(X \geq t) = P(t1_{X\geq t} \geq t) = \mathbb{E}(1_{X\geq t}) \leq \mathbb{E}(\frac{X}{t}) = \frac{\mathbb{E}(X)}{t}.$$

(c) Chernoff gives $e^{50} \cdot (\frac{2}{3})^{150} = (\frac{8e}{27})^{50}$. It is not necessary to simplify this further.

Exercise 3. (/12 P.)

Consider the linear regression problem

$$y = X\beta + \varepsilon$$

with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$.

Suppose we have an orthogonal design matrix, i.e. $X'X = I_{p\times p}$.

(a) Write down the classical ordinary least squares estimator under the assumption of an orthogonal design matrix and denote each component of the vector as $\widehat{\beta}_i^{OLS}$, $i = 1, \ldots, p$. (/4 P.)

(b) Then, the ridge regression problem can be written as

$$\sum_{i=1}^{p}(\beta_i - \widehat{\beta}_i^{OLS})^2 + \lambda \sum_{i=1}^{p} \beta_i^2.$$

Derive the ridge regression estimator for the ith component in terms of $\widehat{\beta}_i^{OLS}$. (/4 P.)

(c) Let X be an $n \times p$ design matrix where $n = 8$ and $p = 10$, representing information about various features of plants. Let $y \in R^n$ be a vector of the plants' size, such that y_i represents the size of the ith plant. We would like to train a regression model on this data. Which method would be a reasonable choice for this task? Explain your choice! (/4 P.)

Solution

(a) Since we know that the OLS estimator is $\hat{\beta} = (X'X)^{-1}X'y$, and since $(X'X)^{-1}$ is the unit matrix, this simplifies to $\hat{\beta} = X'y$.

(b) This means, that the ridge regression estimator simplifies to (same argument as before...)

$$\widehat{\beta}^{RIDGE} = \frac{1}{\lambda + 1} X'y.$$

(c) Since we have more features than data points, it is quite likely that we suffer overfitting; hence we prefer a method that is capable of feature selection. This is LASSO.

Exercise 4. (/12 P.)

(a) Let $X = (X_1, X_2, X_3, X_4, X_5)$ be a random vector distributed as $X \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma^{-1} = \begin{pmatrix} 3 & 0 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix}$$

a) What is the graph for X, viewed as an undirected graphical model? (/4 P.)

b) Which of the following independence statements are true? (No explanation necessary.)

i. $X_1 \perp\!\!\!\perp X_2 | X_3$

ii. $X_1 \perp\!\!\!\perp X_5 | X_2, X_3, X_4$

iii. $X_1 \perp\!\!\!\perp X_3 | X_2, X_4, X_5$

iv. $X_1 \perp\!\!\!\perp X_5 | X_3$ (/4 P.)

(b) Let $Y_1, Y_2, Y_3 \sim \mathcal{N}(0, 1)$ be independent. Define

$$X_1 = Y_1, \quad X_2 = aX_1 + Y_2, \quad X_3 = bX_2 + cX_1 + Y_3$$

where a, b, c are nonzero constants. Now suppose that $c = -b\frac{a^2+1}{a}$. Show that

$$\text{Cov}(X_2, X_3) = 0.$$

(/4 P.)

Solution

(a) The graph contains edge $(1, 3)$, and the path $(2, 3, 4, 5)$.

a) $X_1 \perp\!\!\!\perp X_2 | X_3$ true (3 blocks paths from 1 to 2)

b) $X_1 \perp\!\!\!\perp X_5 | X_2, X_3, X_4$ true $(\Sigma^{-1})_{15} = 0$

c) $X_1 \perp\!\!\!\perp X_3 | X_2, X_4, X_5$ false $(\Sigma^{-1})_{13} \neq 0$

d) $X_1 \perp\!\!\!\perp X_5 | X_3$ true (3 blocks paths from 1 to 5)

(b) For zero-mean random variables, Cov is a bi-linear function, hence we can treat it like a product:

$$\begin{aligned} &Cov(aY_1 + Y_2, abY_1 + bY_2 - b(\frac{a^2+1}{a})Y_1 + Y_3) \\ = &a^2bCov(Y_1, Y_1) - b(a^2+1)Cov(Y_1, Y_1) + bCov(Y_2, Y_2) \\ &= a^2b - a^2b - b + b = 0. \end{aligned}$$

We note that we have omitted all $Cov(Y_i, Y_j)$-terms with $i \neq j$, evaluating to 0.