EPFL

Profs. Martin Jaggi and Rüdiger Urbanke
Machine Learning – CS-433 - IN
Wednesday 15.01.2020
from 16h15 to 19h15 in STCC08328
Duration : 180 minutes

1

Heidi

SCIPER : **837495**

Do not turn the page before the start of the exam. This document is double-sided, has 16 pages. Do not unstaple.

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet (hand-written or 11pt min font size) if you have one; place all other personal items below your desk.
- You each have a different exam.
- Only answers in this booklet count. No extra loose answer sheets. Use loose sheets as scrap paper.
- For the **multiple choice** questions, we give :
  +2 points if your answer is correct,
  0 points if you give no answer or more than one,
  −0.5 points if your answer is incorrect.
- For the **true/false** questions, we give :
  +1 points if your answer is correct,
  0 points if you give no answer or more than one,
  −1 points if your answer is incorrect.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question turns out to be wrong or ambiguous, we may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer
Antwort auswählen

ne PAS choisir une réponse | NOT select an answer
NICHT Antwort auswählen

Corriger une réponse | Correct an answer
Antwort korrigieren

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

First part: multiple choice questions

For each question, mark the box corresponding to the correct answer. Each question has **exactly one** correct answer.

Shrinking Confidence

Assume that you get a confidence interval of size δ for some problem given N iid samples.

Question 1 Expressed as a function of N, how many iid samples do you need to get a confidence interval of size $\delta/3$?

- [ ] $3N$
- [ ] $N/3$
- [ ] N^3
- [x] $9N$
- [ ] $\sqrt{3N}$
- [x] e^{3N}

conf. interval decay $= \frac{1}{\sqrt{N}} = \frac{1}{3}$

$\sqrt{N} = 3$
$N = 9$

Family Expectations

For $\theta > 0$, consider the following probability distribution defined for $y \geq 0$,

$$p(y) = e^{(-y^3)\theta - A(\theta)}, \quad A(\theta) = -\frac{1}{3}\log(\theta) + c,$$

where c is an appropriate constant.

Question 2 What is $\mathbb{E}_{Y\sim p}[Y^3]$, expressed as a function of θ?

- [x] θ^3
- [ ] y
- [ ] $A(\theta)$
- [ ] c
- [ ] 1
- [x] $\frac{1}{3\theta}$

???

$A'(\theta) = -\frac{1}{3\theta}$

$A''(\theta) = \frac{1}{3\theta^2}$

$A''' = -\frac{2}{3\theta^3}$

$A'(\theta) = \mathbb{E}[\phi(y)]$

SVMs versus Logistic Regression

Consider a classification problem using either SVMs or logistic regression and separable data. For logistic regression we use a small regularization term (penality on weights) in order to make the optimum well-defined. Consider a point that is correctly classified and distant from the decision boundary. Assume that we move this point slightly.

Question 3 What will happen to the decision boundary?

- [ ] Small change for SVMs and small change for logistic regression.
- [ ] No change for SVMs and large change for logistic regression.
- [ ] No change for SVMs and no change for logistic regression.
- [x] No change for SVMs and a small change for logistic regression.
- [ ] Large change for SVMs and large change for logistic regression.
- [ ] Large change for SVMs and no change for logistic regression.
- [ ] Small change for SVMs and no change for logistic regression.
- [ ] Small change for SVMs and large change for logistic regression.
- [ ] Large change for SVMs and small change for logistic regression.

KNN Classifier

You are in D-dimensional space and use a KNN classifier with $k = 1$. You are given N samples and by running experiments you see that for most random inputs $\mathbf{x}$ you find a nearest sample at distance roughly δ.

Question 4 You would like to decrease this distance to $\delta/2$. How many samples will you likely need? Give an educated guess.

- [x] $2^D N$
- [ ] N^D
- [ ] $2D$
- [ ] $\log(D)N$
- [ ] N^2
- [x] D^2
- [ ] $2N$
- [ ] DN

Volume of ball $\boxed{r^D}$

$N \cdot r^D$, $r = 2$.

Tricky Question

You are given samples $\mathcal{S} = \{\mathbf{x}_n\}_{n=1}^N$, where each sample has two components, i.e., $\mathbf{x} = (x_1, x_2)$. You compute from this the corresponding kernel matrix $\mathbf{K}$ with entries $\mathbf{K}_{i,j} = \mathbf{x}_i^T \mathbf{x}_j$.

Assume now that you transform the feature vector to $\tilde{\mathbf{x}} = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$ and compute from this new feature vector the corresponding kernel matrix $\tilde{\mathbf{K}}$ with entries $\tilde{\mathbf{K}}_{i,j} = \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_j$.

Question 5 What function does this transform correspond to? I.e., what function $f(\cdot)$ can you pick so that $\tilde{\mathbf{K}} = f(\mathbf{K})$, where the function $f(\cdot)$ is applied component-wise?

- ☐ $f(z) = e^z$
- ☑ $f(z) = z^2$
- ☐ $f(z) = 1$
- ☐ $f(z) = z^3$
- ☐ $f(z) = \log(z)$
- ☐ $f(z) = z$
- ☐ $f(z) = \sqrt{2} z$

Houston we have an Overflow Problem

Assume that you want to implement the sigmoid function: $\sigma(x) = e^x/(e^x + 1)$. You know that your computer can handle numbers with very small absolute value but might overflow when dealing with numbers that have a very large absolute value. Let $f_1(x) = e^x/(e^x + 1)$ and $f_2(x) = 1/(e^{-x} + 1)$.

Question 6 Which of the following implementations is best?

- ☐ $f_1(x)$
- ☑ $f_2(x)$
- ☐ $f_1(x)$ if $x > 0$ and $f_2(x)$ otherwise
- ☐ ? $f_2(x)$ if $x > 0$ and $f_1(x)$ otherwise

K-means Clustering

Consider K-means clustering in D-dimensional real space and assume that K is known. We have N samples. We have seen in class that this corresponds to solving the following optimization problem:

$$\min_{\mathbf{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z},\boldsymbol{\mu}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$
$$\text{where } \mathbf{z}_n = [z_{n1}, z_{n2}, \dots, z_{nK}]^\top$$
$$\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]^\top$$
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K]^\top$$

Question 7 What extra conditions do we need to render this formulation correct?

- [ ] s.t. $\boldsymbol{\mu}_k \in \mathbb{R}^N$, $z_{nk} \in [0,1]$, $\sum_{n=1}^{N} z_{nk} = 1$
- [x] s.t. $\boldsymbol{\mu}_k \in \mathbb{R}^D$, $z_{nk} \in \{0,1\}$, $\sum_{k=1}^{K} z_{nk} = 1$
- [ ] s.t. $\boldsymbol{\mu}_k \in \mathbb{R}^N$, $z_{nk} \in \{0,1\}$, $\sum_{n=1}^{N} z_{nk} = 0$
- [ ] s.t. $\boldsymbol{\mu}_k \in \mathbb{R}^D$, $z_{nk} \in \{-1,1\}$, $\sum_{k=1}^{K} z_{nk} = 0$
- [ ] s.t. $z_{nk} \in \mathbb{R}^D$, $\boldsymbol{\mu}_k \in \{-1,1\}$, $\sum_{k=1}^{K} z_{nk} = 1$
- [ ] s.t. $z_{nk} \in \mathbb{R}^D$, $\boldsymbol{\mu}_k \in \{0,1\}$, $\sum_{k=1}^{K} z_{nk} = 1$
- [ ] s.t. $z_{nk} \in \mathbb{R}^D$, $\boldsymbol{\mu}_k \in \{0,1\}$, $\sum_{k=1}^{K} \boldsymbol{\mu} = 1$
- [ ] s.t. $z_{nk} \in \mathbb{R}^K$, $\boldsymbol{\mu}_k \in \{-1,1\}$, $\sum_{k=1}^{K} z_{nk} = 0$
- [ ] s.t. $\boldsymbol{\mu}_k \in \mathbb{R}^K$, $z_{nk} \in [0,1]$, $\sum_{n=1}^{N} z_{nk} = 0$

Finding Adversarial Examples

Consider a binary classification problem with classifier $f(\mathbf{x})$ given by

$$f(\mathbf{x}) = \begin{cases} 1, & g(\mathbf{x}) \geq 0, \\ -1, & g(\mathbf{x}) < 0, \end{cases}$$

and $\mathbf{x} \in \mathbb{R}^6$. Consider a specific pair $(\mathbf{x}, y = 1)$ and assume that $g(\mathbf{x}) = 8$. In particular this means that this point is classified correctly by f. Assume further that we have computed the gradient of g at $\mathbf{x}$ to be $\nabla_{\mathbf{x}} g(\mathbf{x}) = (+1, -2, +3, -4, +5, -6)$. You are allowed to make one step in order to (hopefully) find an adversarial example. In the following four questions, assume $\epsilon = 1$.

Question 8 Which offset δ with $\|\delta\|_1 \leq 1$ yields the smallest value for $g(\mathbf{x} + \delta)$, assuming that g is (locally) linear?

- [x] $(0, 0, 0, 0, 0, 1)$
- [ ] $(+1, -1, +1, -1, +1, -1)$
- [ ] $(+1, -2, +3, -4, +5, -6)$
- [ ] $(+1, +1, +1, +1, +1, +1)$
- [ ] $(-1, +2, -3, +4, -5, +6)$
- [ ] $-(0, 0, 0, 0, 0, 1)$
- [ ] $(-1, +1, -1, +1, -1, +1)$
- [ ] $(-1, -1, -1, -1, -1, -1)$

Question 9 What is the value of $g(\mathbf{x} + \delta)$ for this ℓ_1-optimal choice assuming that g is (locally) linear?

- [ ] $+13$
- [ ] -4
- [ ] -5
- [ ] -7
- [x] 2
- [ ] 4
- [ ] -13
- [ ] -2
- [ ] $+7$
- [ ] 0

$g(x+\delta) = x + \delta^T \nabla_x g(x)$

$\|\delta\| = 6 \;;\; (-6)$

$g(x) = 8 \quad g(\delta) = \delta^T \nabla_x g(x)$

$g(x-\delta) = 8-6 = 2$

Question 10 Which offset δ with $\|\delta\|_\infty \leq 1$ yields the smallest value for $g(\mathbf{x}+\delta)$, assuming that g is (locally) linear?

- ☐ $(+1, -2, +3, -4, +5, -6)$
- ☐ $-(0, 0, 0, 0, 0, 1)$
- ☐ $(0, 0, 0, 0, 0, 1)$
- ☐ $(-1, -1, -1, -1, -1, -1)$
- ☐ $(+1, +1, +1, +1, +1, +1)$
- ☐ $(-1, +1, -1, +1, -1, +1)$
- ☐ $(+1, -1, +1, -1, +1, -1)$
- ☐ $(-1, +2, -3, +4, -5, +6)$

Question 11 What is the value of $g(\mathbf{x}+\delta)$ for this ℓ_∞-optimal choice assuming that g is (locally) linear?

- ☐ -5
- ☐ -2
- ☐ -7
- ☐ $+7$
- ☐ 4
- ☐ 0
- ☐ $+13$
- ☐ 2
- ☐ -4
- ☐ -13

Second part: true/false questions

For each question, mark the box (without erasing) TRUE if the statement is **always true** and the box FALSE if it is **not always true** (i.e., it is sometimes false).

Question 12 (Bayes Nets) We are given a Bayes net involving the variables $X_1, \cdots, X_n$. We determine, using our standard rules, that $X_1 \perp X_2 \mid X_3$.
Assume now that you delete some edges in the original Bayes net. For the modified Bayes net, is it *always* true that $X_1 \perp X_2 \mid X_3$?

☐ TRUE ☒ FALSE

Question 13 (Nearest Neighbor) The training error of the 1-nearest neighbor classifier is zero.

☒ TRUE ☐ FALSE

Question 14 (Backpropagation) Training via the backpropagation algorithm always learns a globally optimal neural network if there is only one hidden layer and we run an infinite number of iterations and decrease the step size appropriately over time.

☐ TRUE ☒ FALSE

Question 15 (Infinite Data) Assume that your training data $\mathcal{S} = \{(\mathbf{x}_n, y_n)\}$ is iid and comes from a fixed distribution $\mathcal{D}$ that is unknown but is known to have bounded support. Assume that your family of models contains a finite number of elements and that you choose the best such element according to the training data. You then evaluate the risk for this chosen model. Call this the training risk. As $|\mathcal{S}|$ tends to infinity, this training risk converges to the true (according to the distribution $\mathcal{D}$) risk of the best model in this family.

☑ TRUE ☐ FALSE

Question 16 (Robustness) The l_1 loss is less sensitive to outliers than l_2.

☑ TRUE ☐ FALSE

Question 17 (Convex I) Unions of convex sets are convex.

☑ TRUE ☐ FALSE

Question 18 (Convex II) Intersections of convex sets are convex.

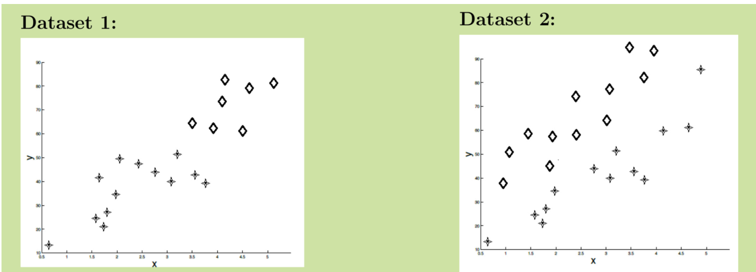
☑ TRUE ☐ FALSE

Question 19 (Convex III) Let $f, g : \mathbb{R} \to \mathbb{R}$ be two convex functions. Then $h = f \circ g$ is always convex.

☑ TRUE ☐ FALSE check case of f' decreasing.

Question 20 (Maximum Likelihood) Assume that $X \in \{0,1\}$ and that $p(X=0) = \frac{1}{3}$. Assume further that $Y = X + Z$ where Z is a zero-mean Gaussian noise of variance 1. We observe Y and are asked to guess X. The maximum likelihood estimator $\hat{X}(Y) = \text{argmax}_{x \in \{0,1\}} p(Y=y|X=x)$ minimizes the probability of error.

☑ TRUE ☐ FALSE



Question 21 (Due to Matt Gormley) (PCA) Consider the two datasets given in the figure. Assume that you first project the points into the first principal component and then you use a threshold function to classify the data. This approach works better for Dataset 2 than for Dataset 1.

☐ TRUE ☒ FALSE

Question 22 (Stochastic Gradient Descent) One iteration of standard SGD for SVM, logistic regression and ridge regression costs roughly $\mathcal{O}(D)$, where D is the dimension of a data point.

☑ TRUE ☐ FALSE

Question 23 (Stochastic Gradient Descent, cont) SGD in typical machine learning problems requires fewer parameter updates to converge than full gradient descent.

☐ TRUE ☑ FALSE

Question 24 (SGD & Matrix Factorization) For optimizing a matrix factorization problem in the recommender systems setting, as the number of observed entries increases, the computational cost of full gradient steps increases, while the cost of an SGD step remains the same.

☑ TRUE ☐ FALSE

Question 25 (Alternating Least Squares & Matrix Factorization) For optimizing a matrix factorization problem in the recommender systems setting, as the number of observed entries increases but all K, N, D are kept constant, the computational cost of the matrix inversion in Alternating Least-Squares increases.

☐ TRUE ☒ FALSE

Question 26 (Text Representation Learning, GloVe) Learning GloVe word vectors is identical to approximating the observed entries of the word/context co-occurence counts by $\mathbf{W}\mathbf{Z}^{\top}$, in the least square sense, if the f_{dn} weights are set to 1 for all observed entries.

☒ TRUE ☐ FALSE

Question 27 (Text Representation Learning, word2vec) An SGD step for learning GloVe word vectors is computationally equally expensive to an SGD step in word2vec, however only GloVe requires memory the size of the co-occurence matrix.

☒ TRUE ☐ FALSE

Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Do not cross any checkboxes, they are reserved for correction.

Independence

Consider the following joint distribution that has the factorization

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_1, x_3)p(x_5|x_4).$$

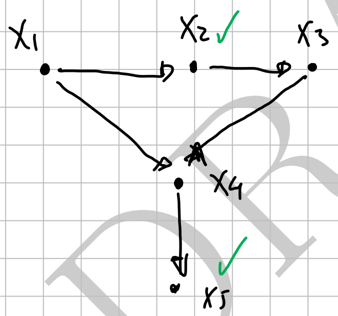
Question 28: (*4 points.*) Determine whether the following statement is correct.

$$X_1 \perp X_3 \mid X_2, X_5$$

Show your reasoning.

☐0 ☐1 ☐2 ☐3 ☐4

This is not correct, as we need to know the probability of X_1 to know



Question 29: (*3 points.*) We say that a data point y follows a Poisson distribution with parameter θ if the probability of the observation y, $y \in \mathbb{N}$, is given by

$$p(y \mid \theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

Assume that you are given the samples $\mathcal{S} = \{y_1, \cdots, y_N\}$.

(a) Write down the log-likelihood, call it $\mathcal{L}$, of these samples as a function of θ assuming that the samples are iid and follow a Poisson distribution with parameter θ.

(b) What is the parameter θ that maximizes this log-likelihood expressed as a function of the samples?

(c) Interpret the result.

☐ 0 ☐ 1 ☐ 2 ☐ 3

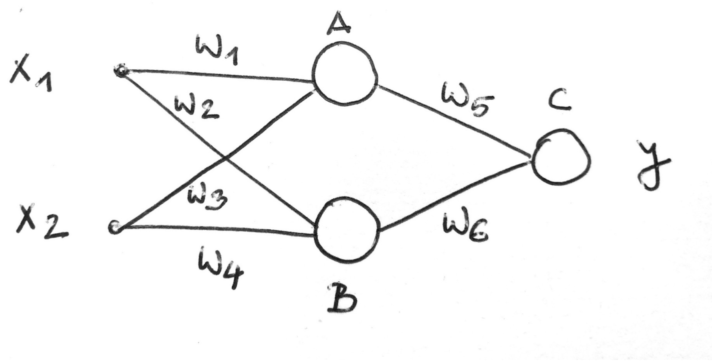
$$\operatorname{argmin}_{\theta} \ell \prod \frac{\theta^y e^{-\theta}}{y!} = \frac{d}{d\theta} \sum y \log\theta - \theta - \lg(y!)$$

$$\sum \left[y \frac{1}{\theta} - 1 \right] = 0$$

$$\frac{\sum y}{\theta} = N$$

$$\theta = \frac{\sum y}{N}$$

DRAFT



Question 30: (*4 points.*)[Question due to Eric Xing and Tom Mitchell] Consider the neural network shown in the figure. This network has one input layer, one hidden layer and one output layer. Further, it has two input features, denote them by x_1 and x_2, and one output, call it y.

Assume that we have two types of general activiation functions.

(a) $S : S(a) = \text{sign}(\frac{1}{1+e^{-a}} - \frac{1}{2})$,

(b) $L : L(a) = \gamma a$, for some scalar γ

where in both cases $a = \sum_i w_i x_i$ denotes the sum over the incoming connections to the current neuron.

(a) Assume that we assign the activation function L to all three nodes A, B, and C. What model does this network express? Explicitly express the function computed by the network in terms of $w_1, ...w_6$.

(b) Assume that we assign the activation function L to nodes A and B and the activation function S to node C. What model does this network express? Does it remind you of something you have seen in class? Explicitly express the function computed by the network in terms of $w_1, ..., w_6$.

Note that we are only interested in the resulting input-output relationship and there is no loss function involved.

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4

1

$A = \gamma (w_1 x_1 + w_3 x_2)$

$C = \gamma (w_5 A + w_6 B)$

$B = \gamma (w_3 x + w_4 x_2)$

$C = \gamma^2 (w_5 (w_1 x_1 + w_3 x_2) + w_6 (w_3 x_1 + w_4 x_2))$

$C = \gamma^2 (x_1 (w_5 w_1 + w_6 w_3) + x_2 (w_5 w_3 + w_6 w_4))$

linear regression.