Ensimag – Grenoble INP – UGA Year 2022-2023
Statistical Analysis and Document Mining
Pedro L. C. Rodrigues pedro.rodrigues@inria.fr
Jacopo Iollo jacopo.iollo@inria.fr

List of multiple choice questions

Each question has **exactly one correct** answer.

– Question 1: One gaussian (credits to Berkeley CS-189)

Consider a d-dimensional multivariate normal distribution that is isotropic (i.e., its isosurfaces are spheres). Let $\mathbf{\Sigma}$ be its $d \times d$ covariance matrix. Let $\mathbf{I}$ be the $d \times d$ identity matrix. Let σ be the standard deviation of any one component (feature). Then

(A) $\mathbf{\Sigma} = \sigma \mathbf{I}$

(B) $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$

(C) $\mathbf{\Sigma} = \frac{1}{\sigma} \mathbf{I}$

(D) None of the above.

– Question 2: Two gaussians (credits to Berkeley CS-189)

Say we have two 2-dimensional Gaussian distributions representing two different classes. Which of the following conditions will result in a linear decision boundary

(A) Same mean for both classes.

(B) Different covariance matrix for each class.

(C) Same covariance matrix for both classes.

(D) None of the above.

– Question 3: Logistic regression (credits to EPFL CS-433)

Consider the logistic regression loss $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ for a binary classification task with data $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(\log\left(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right) - y_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)$$

Which of the following is a gradient of the loss $\mathcal{L}$?

(A) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} - y_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)$

(B) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)$

(C) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \left(\frac{1}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} - y_i \right)$

(D) None of the above.

– Question 4: Gaussian discriminant analysis (credits to Berkeley CS-189)

Which of the following are true about two-class Gaussian discriminant analysis? Assume you have estimated the parameters $\hat{\boldsymbol{\mu}}_1, \hat{\mathbf{\Sigma}}_1, \hat{\pi}_1$ for class 1 and $\hat{\boldsymbol{\mu}}_0, \hat{\mathbf{\Sigma}}_0, \hat{\pi}_0$ for class 0.

(A) If $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0$ and $\hat{\pi}_1 = \hat{\pi}_0$, then the LDA and QDA classifiers are identical.

(B) If $\hat{\mathbf{\Sigma}}_1 = \boldsymbol{I}$ and $\hat{\mathbf{\Sigma}}_0 = 5\boldsymbol{I}$, then the LDA and QDA classifiers are identical.

(C) If $\hat{\mathbf{\Sigma}}_1 = \hat{\mathbf{\Sigma}}_0$, $\hat{\pi}_0 = 1/6$ and $\hat{\pi}_0 = 5/6$, then the LDA and QDA classifiers are identical.

(D) None of the above.

– **Question 5: Detecting dogs (credits to Berkeley CS-189)**

You want to train a dog identifier with Gaussian discriminant analysis. Your classifier takes an image vector as its input and outputs 1 if it thinks it is a dog, and 0 otherwise. You use the CIFAR10 dataset, modified so all the classes that are not "dog" have the label 0. Your training set has 5k dog images and 45k non-dog ("other") images. Which of the following statements seem likely to be correct.

(A) LDA has an advantage over QDA because the two classes have different numbers of training examples.

(B) QDA has an advantage over LDA because the two classes have different numbers of training examples.

(C) LDA has an advantage over QDA because the two classes are expected to have very different covariance matrices.

(D) QDA has an advantage over LDA because the two classes are expected to have very different covariance matrices.

– **Question 6 (credits to Berkeley CS-189)**

Which of the following are true for the k-nearest neighbor (k-NN) algorithm?

(A) k-NN can be used for both classification and regression

(B) The decision boundary looks smoother with smaller values of k.

(C) As k increases, the variance usually increases

(D) None of the above.

– **Question 7 (credits to Berkeley CS-189)**

Logistic regression

(A) Assumes that the distribution of the data is Gaussian

(B) Has a closed-form solution

(C) Minimizes a convex cost function

(D) None of the above

– **Question 8 (credits to Berkeley CS-189)**

Cross-validation

(A) Is often used to select hyperparameters.

(B) Does nothing to prevent overfitting.

(C) Is guaranteed to prevent overfitting.

(D) None of the above.

– **Question 9: (credits to Berkeley CS-189)**

Good practices to avoid overfitting include

(A) Using a two part cost function which includes a regularizer to penalize model complexity.

(B) Using a good optimizer to minimize error on training data.

(C) Discarding 50

(D) None of the above.

– **Question 10: (credits to Berkeley CS-189)**

In the following statements, the word “bias” is referring to the bias-variance decomposition.

(A) A model trained with N training points is likely to have lower variance than a model trained with $2N$ training points.

(B) If my model is underfitting, it is more likely to have high bias than high variance.

(C) Increasing the number of parameters (weights) in a model usually improves the test set accuracy.

(D) None of the above.

– **Question 11: Linear regression (credits to EPFL CS-433)**

Assume we are doing linear regression with mean-squared loss and L2-regularization on four one-dimensional data points. Our prediction model can be written as $f(x) = ax + b$ and the optimization problem can be written as

$$a^\star, b^\star = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{4} \Big(y_i - f(x_i)\Big)^2 + \lambda a^2$$

Assume that our data points (x_i, y_i) are $\{(-2,1), (-1,3), (0,2), (3,4)\}$. What is the optimal value for the bias, $b^\star$?

(A) Depends on the value of λ.

(B) 3

(C) 2.5

(D) None of the above answers.

– **Question 12: Linear regression (credits to EPFL CS-433)**

Under certain conditions, maximizing the log-likelihood is equivalent to minimizing mean-squared error for linear regression. The mean-squared error can be defined as

$$\mathcal{L}_{\text{mse}}(\boldsymbol{\beta}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2$$

and

$$y_i = \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

is assumed for the probabilistic model. Which of following conditions is necessary for the equivalence?

(A) The noise parameter ε_i should have a normal distribution.

(B) The target variable y_i should have a normal distribution.

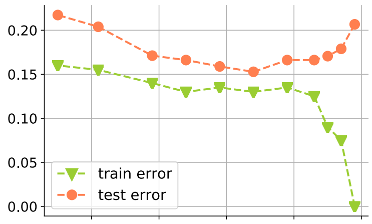
(C) The i.i.d. assumption on the variable $\boldsymbol{\beta}$.

(D) The noise parameter ε_i should have non-zero mean.

– **Question 13: Train/test error**

The above figure was produced by changing a hyperparameter in a classifier on a non-linearly separable training dataset. For which classifier could this picture be produced and how was its hyperparameter changed?

(A) Logistic regression, increasing regularization parameter λ



(B) Logistic regression, decreasing regularization parameter λ

(C) k-nearest neighbor classifier, decreasing number of neighbors k.

(D) k-nearest neighbor classifier, increasing number of neighbors k.

– Question 14: Ridge regression (credits to Berkeley CS-189)

How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?

(A) Ridge has larger bias, larger variance.

(B) Ridge has larger bias, smaller variance.

(C) Ridge has smaller bias, larger variance.

(D) Ridge has smaller bias, smaller variance.

– Question 15: Ridge regression (credits to Berkeley CS-189)

Given a design matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, labels $\mathbf{y} \in \mathbb{R}^N$, and $\lambda > 0$, we find the weight vector $\boldsymbol{\beta}^\star$ that minimizes $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$. Suppose that $\boldsymbol{\beta}^\star \neq \mathbf{0}$.

(A) The variance of the method decreases if λ increases enough.

(B) There may be multiple solutions for $\boldsymbol{\beta}^\star$

(C) The bias of the method decreases if λ increases enough

(D) None of the above.

– Question 16: Principal component analysis (credits to Berkeley CS-189)

Given d-dimensional data $\mathbf{x}_{i=1}^N$, you run principle component analysis and pick P principle components. Can you always reconstruct any data point $\mathbf{x}_i$ for $i \in \{1...N\}$ from the d principle components with zero reconstruction error?

(A) Yes, if $P < d$.

(B) Yes, if $P = d$.

(C) Yes, if $P < n$.

(D) No, always.

– Question 17: Logistic regression assumptions (credits to EPFL CS-433)

Binary logistic regression assumes

(A) Linear relationship between the input variables.

(B) Linear relationship between the observations.

(C) Linear relationship between the input variables and the inverse sigmoid of the probability of the event that the outcome $Y = 1$.

(D) Linear relationship between the input variables and the probability of the event that the outcome $Y = 1$.

– Question 18: Robustness to outliers (credits to EPFL CS-433)

We consider a classification problem on linearly separable data. Our dataset had an outlier – a point that is very far from the other datapoints in distance. We trained the linear discriminant analysis (LDA), logistic regression and 1-nearest-neighbour classifiers on this dataset. We tested trained models on a test set that comes from the same distribution as training set, but doesn't have any outlier points. After that we removed the outlier and retrained our models.

After retraining, which classifier will **not change** its decision boundary around the test points.

(A) Logistic regression.

(B) 1-nearest-neighbors classifier.

(C) LDA.

(D) None of them.

– Question 19: Bias-variance decomposition (credits to EPFL CS-433)

Consider a regression model where data (x, y) is generated by input $x \in \mathbf{R}$ uniformly sampled between $[0, 1]$ and $y = x + \varepsilon$, where ε is random noise with mean 0 and variance 1. Two models are carried out for regression: model $\mathcal{A}$ is a trained quadratic function $g_{\mathcal{A}}(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$ and model $\mathcal{B}$ is a constant function $g_{\mathcal{B}}(x) = \frac{1}{2}$. Compared to model $\mathcal{B}$, model $\mathcal{A}$ has

(A) Higher bias, higher variance.

(B) Lower bias, higher variance.

(C) Higher bias, lower variance.

(D) Lower bias, lower variance.

– Question 20: PCA (credits to EPFL CS-433)

Which of the following transformations to a data matrix $\mathbf{X}$ will affect the principal components obtained through PCA?

(A) Adding a constant value to all elements of $\mathbf{X}$.

(B) Multiplying one of the features of $\mathbf{X}$ by a constant.

(C) Adding an extra feature to $\mathbf{X}$ (i.e. an extra column) that is constant across all data points.

(D) None of the above answers.

– Question 21: Ridge regularization (credits to EPFL CS-433)

Assume we have N training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ where each $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

For $\lambda \geq 0$, we consider the following loss function:

$$\mathcal{L}_\lambda(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2$$

and let $C_\lambda = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_\lambda(\boldsymbol{\beta})$ denote the optimal loss value. Which of the following statements is **true**?

(A) C_λ is a non-increasing function of λ.

(B) For $\lambda = 0$, the loss $\mathcal{L}_0$ is non-convex and might have several minimizers.

(C) C_λ is a non-decreasing function of λ.

(D) None of the above statements are true.

– Question 22: Linear regression (credits to Berkeley CS-189)

In linear regression, we model $p(y \mid \mathbf{x}) \sim \mathcal{N}(\beta^\top \mathbf{x} + \beta_0, \sigma^2)$. The irreducible error in this model is

(A) σ^2

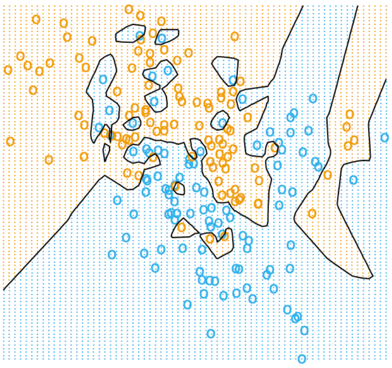
(B) $\mathbb{E}[y \mid \mathbf{x}]$

(C) $\mathbb{E}[(y - \mathbb{E}[y \mid \mathbf{x}])^2 \mid \mathbf{x}]$

(D) None of the above.

– Question 23: Classifier boundary (credits to EPFL CS-189)

Which of these classifiers could have generated the decision boundary here below



(A) Logistic regression

(B) 1-NN

(C) Quadratic discriminant analysis

(D) None of the above.