Homework #2, EECS 553, F26. Due **Fri. Sep. 18**, by 11:00PM

Notes

- Further topics from 501,551, as well as supervised classification methods.

Problems

1. [6] **Gaussian level sets** (501&551)

Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \in \mathbb{R}^{2\times 2}$, where $\boldsymbol{\Lambda} = \mathsf{Diag}\{8, 2\}$ and $\boldsymbol{U}$ is the **orthogonal matrix**:

$$\boldsymbol{U} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

for $\theta = \pi/4$. Suppose that $\mathsf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = [1\ 1]^\top$.

(a) [3] Sketch the boundary of

$$\mathcal{C} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^2 \ : \ \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \le 1 \right\}.$$

Hint. If we change ≤ 1 to $= 1$, then the set $\mathcal{C}$ is an **ellipse**. Your sketch should indicate the center, lengths of major and minor axes, and angle between the major axis and x-axis. Sketches may be hand drawn.

(b) [3] What is $\mathsf{P}\{\mathsf{X} \in \mathcal{C}\}$?
Hint. Use the **chi-squared distribution** and the following property of **multivariate Gaussians**: If $\mathsf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ is a matrix such that the matrix-vector product $\boldsymbol{A}\mathsf{X}$ is defined, then $\boldsymbol{A}\mathsf{X} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$. You should get a value between 0.6 and 0.7.

This problem helps explain the meaning of the elliptical contours shown in the LDA and QDA demos/slides.

2. [9] **PSD matrices**

A matrix $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ is said to be **co-positive** iff for all $\boldsymbol{x} \ge \boldsymbol{0}$ (entrywise non-negative), $\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x} \ge 0$. The verification of matrix copositivity is a well known computationally hard problem, but there are many sufficient conditions that are easier to check.

(a) [3] Prove that a matrix is copositive if it can be decomposed as a sum of a **positive semidefinite** (PSD, denoted by "$\succeq \boldsymbol{0}$") and an element-wise nonnegative matrix:

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} \text{ where } \boldsymbol{A} \succeq 0,\ b_{ij} \ge 0, \quad \forall i, j = 1, \ldots, d.$$

Show that the following types of matrices are PSD:

(b) [3] **Hessian matrices**, *i.e.*, $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$, of a **convex** and twice-differentiable function $f(\boldsymbol{x})$.
Hint. Use the Lagrange form of the remainder in Taylor's theorem for twice-differentiable functions: there exists $t \in (0, 1)$ such that

$$f(\boldsymbol{x} + \boldsymbol{d}) = f(\boldsymbol{x}) + (\nabla f(\boldsymbol{x}))^\top \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x} + t\boldsymbol{d})\boldsymbol{d}, \quad \forall \boldsymbol{x}, \boldsymbol{d}.$$

Also use the following inequality that holds for any differentible convex function f:

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

(c) [3] **Gram matrices**, *i.e.*, any $n \times n$ matrix whose (i, j) entry is $\langle \boldsymbol{x}_i,\ \boldsymbol{x}_j \rangle$ for some vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

3. [6] **Maximum-likelihood estimation**

Consider a random variable X (possibly a vector) whose distribution (density function or mass function) belongs to a parametric family. The density or mass function may be written $\mathsf{p}(\boldsymbol{x}\,;\,\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is called the parameter, and can be either a scalar or vector. For example, in the Gaussian family, $\boldsymbol{\theta}$ can be a two-dimensional vector consisting of the mean and variance. Suppose the parametric family is known, but the value of the parameter is unknown. It is often of interest to estimate this parameter from an observation of X. **Maximum-likelihood estimation** is one of the most important parameter estimation techniques.

Let $\mathsf{X}_1, \ldots, \mathsf{X}_n$ be iid (independent and identically distributed) random variables distributed according to $\mathsf{p}(\boldsymbol{x}\,;\,\boldsymbol{\theta})$. By independence, the joint distribution of the observations is the product

$$\prod_{i=1}^{n} \mathsf{p}(\boldsymbol{x}_i\,;\,\boldsymbol{\theta})\,.$$

Viewed as a function of $\boldsymbol{\theta}$, this quantity is called the **likelihood** of $\boldsymbol{\theta}$. Because many common pdfs/pmfs have an exponential form, it is often more convenient to work with the **log-likelihood**:

$$\sum_{i=1}^{n} \log \mathsf{p}(\boldsymbol{x}_i\,;\,\boldsymbol{\theta})\,.$$

A **maximum-likelihood estimator** (MLE) of $\boldsymbol{\theta}$ is a function $\hat{\boldsymbol{\theta}}$ of the observed values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ satisfying

$$\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \mathsf{p}(\boldsymbol{x}_i\,;\,\boldsymbol{\theta}),$$

where "arg max" denotes the set of all values achieving the maximum. If there is always a unique maximizer, $\hat{\boldsymbol{\theta}}$ is called *the* maximum likelihood estimate.

The **negative binomial distribution** models the number of failures in a sequence of iid Bernoulli trials before r successes take place. In particular, suppose $\{y_i\}_{i=1}^{m}$ is a sequence of random variables taking values 0 and 1, drawn iid from Bernoulli(θ), with "success probability" θ, where 1 corresponds to "success" and 0 corresponds to "failure". The length m of the sequence is exactly the number of trials required to draw r successes, i.e., we keep drawing a sample from Bernoulli(θ) until we obtain r successes and the number of draws it takes is m. So $m - r$ is the number of failures *before* the rth success, which is a negative binomial random variable, and we write $(m - r) \overset{\text{iid}}{\sim} \text{NB}(r, \theta)$. Note that the θ is the same in NB(r, θ) and Bernoulli(θ), because the negative binomial distribution implies an underlying Bernoulli distribution.

The negative binomial distribution can be used to answer questions like: If I repeatedly flip a coin, what is the probability that the 5th head occurs on the 12th flip?

(a) [3] Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{NB}(r, \theta)$, where NB$(r, \theta)$ refers to the negative binomial distribution. Assuming r is known, determine the maximum-likelihood estimator of θ as a function of observations (realizations) $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$.

(b) [3] Calculate the Hessian of the log-likelihood function and verify that the estimator you found is indeed the *unique* maximizer of the log-likelihood function.

4. [6] **k-NN classifier distance metric**

The conventional nearest-neighbor classifier uses the Euclidean distance $\|\boldsymbol{x} - \boldsymbol{x}_i\|_2$ to define "nearest" for $\boldsymbol{x}_i \in \mathbb{R}^d$. There are many alternatives, some of which are based on transforming the data into a new coordinate system. For each part below, discuss briefly whether the described approach leads to improved, worse, or the same classification accuracy, or if there is insufficient information to know. *Ignore finite numerical precision effects.*

(a) [2] The discrete wavelet transform (**DWT**) is a classical signal processing method for transforming signals into a domain that is more compressible. (It is used in the JPEG 2000 image compression standard.) A wavelet enthusiast decides to transform all training and test data using $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{W}\boldsymbol{x}$ where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is a DWT matrix. This researcher then uses Euclidean distance in the wavelet coefficient domain, *i.e.*, $\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_i\|_2$, to define "nearest" for a k-NN classifier.
Discuss. Hint. The only thing you need to know about $\boldsymbol{W}$ here is that it is an orthogonal matrix; the rest of the text is just context so that you realize is a plausible approach.

(b) [2] A different researcher uses the transformation $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{Q}\boldsymbol{x}$ where $\boldsymbol{Q}$ is a semi-unitary matrix (orthonormal columns). (This scenario corresponds to an over-complete frame, another popular signal processing tool, and one that has been used to analyze deep networks.) Discuss.

(c) [2] The Julia k-NN demo uses the transformation $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{U}^\top \boldsymbol{x}$ where $\boldsymbol{U}$ is semi-unitary. Discuss.

5. [8] **k-NN classification of digits**

Extend the MNIST k-NN demo to apply k-NN classification in the original "ambient" space of the 28×27 images. Use the same set of train/validate/test cases to compare k-NN in the $d = 2$ embedded PCA feature space to the $d = 28 \cdot 27 = 756$ pixel space.

(a) [3] Submit a screen shot of your code

(b) [5] Plot error vs k like the one in the demo but with 6 curves: 3 curves for the $d = 2$ case, and 3 curves for $d = 756$.

You may use JULIA or Python for this work and it is fine to use AI tools to help translate the JULIA code to Python code if you think that will be helpful.

6. [11] **LDA binary classifier**

The LDA demo classifies samples of MNIST digits "4" and "7" using a "brute-force" arg max approach. Modify it so that it uses the linear decision boundary derived analytically in class. (You may use JULIA or Python for this work and it is fine to use AI tools to help translate the JULIA code to Python code if you think that will be helpful.)

(a) [2] Report the values of $\boldsymbol{w}$ and b.

(b) [9] Submit a 2D plot similar to this example in the demo that shows the training data and the decision regions.
- (2) Submit your plot and a screenshot of your added code to gradescope.
- (2) Report the *test* error rate in the title of the plot.
- (5) Superimpose on the plot the analytical decision boundary corresponding to $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$.

7. [5] **Make your own problem**

Make a clearly stated problem that could be used on an EECS 553 exam based on the topics covered by the due date of this HW set, and provide your own solution to the problem.

A well-designed problem (not too trivial, not too hard) might be used on an actual exam. To earn full credit:

- Create your problem in `hw2.tex` in the Overleaf project linked below.

  An Overleaf tutorial can be found here.
- After you complete editing your question in the shared .tex file, click the "Recompile" button and make sure there is no error in your code. No credit will be earned if the compiling result indicates there are errors in your LaTeX code.
- Your problem must be written using LaTeX using the Overleaf link on EECS 553 Canvas home page.

  Do not share that link with outsiders or turn off "Track changes" in the Overleaf project.
- We strongly encourage you to first make a copy of the template and draft your question and solution there. Then test the LaTeX compiling in your copy. Then copy and paste your part into the shared template.
- Your problem must be well-defined and the solution to the problem must be correct.
- Your problem should not be excessively trivial (nor beyond the scope of the course).
- Your problem must not be identical to any homework or clicker problem.
- Your problem must be in *multiple choice* form and must not include words like "prove that ...".
- You must strictly follow the provide template, i.e., put your unique name in `\newquestion{}`, and type all the answer choices to your question using `\item` under the `enumerate` command; then type the solution within `\answer{}`.
- Do not include other LaTeX packages or define new macros. If there is a macro or package you would like us to include, you can suggest it in Piazza for consideration. You may use the numerous useful macros in `math_commands.tex` to save typing, like `\mA` for matrix $\boldsymbol{A}$ and `\vx` for vector $\boldsymbol{x}$.
- To insert math equations, use $ ... $ for inline math and $$ ... $$ for displayed math.
- Do not include figures, tables, or citations.
- Do not use cross-references, i.e., commands such as `\label{}`, `\ref{}`, `\eqref{}`.
- When adding your question, put it below the existing questions. It is a violation of the Honor Code to tamper with other students' questions.
- Feel free to ask for LaTeX help in Piazza too!

Non-graded problems

8. [4] **Normalization constant for a multidimensional Gaussian** (501&551)

Prove that the normalization constant for a d-dimensional Gaussian distribution is given by

$$|2\pi\boldsymbol{\Sigma}|^{1/2} = \int_{\mathbb{R}^d} \mathrm{e}^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \, \mathrm{d}\boldsymbol{x} \,.$$

Hint. Diagonalize $\boldsymbol{\Sigma}$ and use the relationship between determinant and eigenvalues to write the joint pdf as a product of d 1D Gaussians in a transformed coordinate system. You may use the (known) normalization constant for *univariate* Gaussians.