EECS 545 – Machine Learning - Homework #1

Due: 11:59PM 09/18/2019

Homework Submission: Homeworks must be submitted via Gradescope (https://gradescope.com/) as pdf files. This includes your code when appropriate. Please use a high quality scanner if possible, as found at the library or your departmental copy room. If you must use your phone, please don't just take photos, at least use an app like CamScanner that provides some correction for shading and projective transformations. The gradescope entry code for this course is 9KNG2E, use this code to add the course to your account. If you have never used gradescope before, please register and familiarize yourself with it well before the homework deadline. Gradescope offers tutorials for students, and the best way to learn is to upload a draft of your solutions well before the deadline.

For these problems it may be helpful to recall some the definitions. The image of a matrix $\mathbf{D}$ contains linear combination of the columns of $\mathbf{D}$ and the nullspace of $\mathbf{D}$ contains linear combinations of any vectors $\{\mathbf{u}_i\}$ that satisfy $\mathbf{D}\mathbf{u}_i = 0$. The column (row) rank of an $m \times n$ matrix $\mathbf{D}$ is defined as the size of any basis that spans its columns (rows). $\mathbf{D}$ is said to be full column (row) rank when the associated rank is equal to n (m). A full rank square $n \times n$ matrix $\mathbf{D}$ is invertible. For a product $\mathbf{DE}$ of two compatible (not necessarily square) matrices $\mathbf{D}$ and $\mathbf{E}$, colrank($\mathbf{DE}$) $\leq$min{colrank($\mathbf{D}$), colrank($\mathbf{E}$)}, and similarly for rowrank. An orthogonal matrix is a square matrix $\mathbf{U}$ that satisfies $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$. An idempotent matrix is a square matrix $\mathbf{D}$ that satisfies $\mathbf{D}^2 = \mathbf{D}$. If $\mathbf{D}$ is both idempotent and symmetric, i.e., $\mathbf{D}^T = \mathbf{D}$, $\mathbf{D}$ is a projection matrix and $\mathbf{Dx}$ is an orthogonal projection of a vector $\mathbf{x}$ onto the subspace spanned by the columns of $\mathbf{D}$.

1) **Orthogonal matrices (15 pts).**

(a) (5 pts) Show that if $\mathbf{U}$ is an orthogonal matrix, then for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\| = \|\mathbf{Ux}\|$, where the norm is the Euclidean norm.

(b) (5 pts) Show that each of the following two 2×2 matrices is an orthogonal matrix[1]

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

for $\theta \in [0, 2\pi)$. Give a geometric interpretation of the effect of these two transformations on vector $[x_1, x_2]^T \in \mathbb{R}^2$.

(c) (5 pts) An ellipse in $\mathbb{R}^2$ can be expressed in the form

$$\left\{\mathbf{x} \,\middle|\, (\mathbf{x} - \mathbf{c})^T \mathbf{A} (\mathbf{x} - \mathbf{c}) = r^2 \right\},$$

where $\mathbf{c} \in \mathbb{R}^2$, $r > 0$, and $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ where $\mathbf{U}$ is orthogonal and $\boldsymbol{\Lambda}$ is diagonal. Choose $\mathbf{c}, r, \mathbf{U}$, and Λ such that the major axis has length 3, the minor axis has length 1, and the center of the ellipse is at the point $[3, -1]^T$, and the major axis makes an angle of $+\pi/6$ radians with the positive x-axis.

[1] In fact any 2×2 orthogonal matrix can be represented this way, but you are not asked to show this.

2) **Projection matrices (15 pts).** Consider the 2×2 matrices

$$\mathbf{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}$$

- **(a)** (5 pts) Show that matrices **A** and **B** are idempotent, find their ranks, specify their column span, and specify a basis for the column span. What about **C**? Is it idempotent? Is it orthogonal? What is its rank and column span?

- **(b)** (5 pts) What is the geometric interpretation of these matrices, i.e., how do they transform a vector $[x_1, x_2]^T \in \mathbb{R}^2$

- **(c)** (5 pts) Find the products **AB**, **AC** and **CA**. Specify the ranks of each product. What do these matrix products mean geometrically, i.e. how do they transform a vector $[x_1, x_2]^T \in \mathbb{R}^2$?

3) **Batch centering and scaling (15 pts).** Let $\mathbb{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be a $d \times n$ data matrix whose columns are instances of feature vectors, with the i-th row representing the i-th feature, and define $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^n$ as a vector of ones.

- **(a)** (5 pts) Show that right multiplying $\mathbb{X}$ by the projection matrix $\mathbf{\Pi} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ produces a matrix $\mathbb{X}_c$ whose column average is equal to zero, i.e., the columns of $\mathbb{X}$ have been centered about the sample mean. Specifically, establish that the mean column vector $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{k=1}^{n} \mathbf{x}_k$ is equal to $\mathbb{X}\mathbf{1}\frac{1}{n}$ and that $\mathbb{X}_c\mathbf{1}\frac{1}{n}$ equals zero.

- **(b)** (5 pts) Establish the following representation of the Gram product between the columns of $\mathbb{X}_c$

$$\mathbb{X}_c\mathbb{X}_c^T = \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

  and thus establish that $(n-1)^{-1}\mathbb{X}_c\mathbb{X}_c^T$ is equal to the $d \times d$ sample covariance matrix, often denoted $\widehat{\mathbf{\Sigma}}_X$, of the columns of $\mathbb{X}$. In particular show that the ij element of $(n-1)^{-1}\mathbb{X}_c\mathbb{X}_c^T$ has the standard form of sample covariance between the i-th and j-th features

$$\hat{\sigma}_{ij}^2 = (n-1)^{-1} \sum_{k=1}^{n} (x_{ik} - \hat{\mu}_i)(x_{jk} - \hat{\mu}_j),$$

  where x_{ik} is the i-th element of $\mathbf{x}_k$ and $\hat{\mu}_i = n^{-1}\sum_{l=1}^{n} x_{il}$ is the mean of the i-th feature.

- **(c)** (5 pts) Use the fact that, for any compatible matrices, $\mathbf{A}(\mathbf{I} - \mathbf{B})\mathbf{C} = \mathbf{AC} - \mathbf{ABC}$ and the projection representation $\mathbb{X}_c = \mathbb{X}\mathbf{\Pi}$ of part (a), to establish the identity

$$\widehat{\mathbf{\Sigma}}_X = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^T - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\right)$$

4) **Probability (20 pts).**

(a) (10 pts) Let random variables X and Y be jointly distributed with distribution $p(x, y)$. You can assume that they are jointly discrete so that $p(x, y)$ is the probability mass function (pmf). Show the following results by using the fundamental properties of probability and random variables.

(i) $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$, where $\mathbb{E}[X] = \sum_x xp(x)$ denotes statistical expectation of X and $\mathbb{E}[X|Y] = \sum_x xp(x|Y)$ denotes conditional expectation of X given Y.

(ii) $\mathbb{E}[I[X \in \mathcal{C}]] = P(X \in \mathcal{C})$, where $I[X \in \mathcal{C}]$ is the indicator function of an arbitrary set $\mathcal{C}$ (i.e. $I[X \in \mathcal{C}] = 1$ if $X \in \mathcal{C}$ and 0 otherwise.

(iii) If X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

(iv) If X and Y take values in $\{0, 1\}$ and $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, then X and Y are independent.

(b) (10 pts) For the following equations, describe the relationship between them. Write one of four answers: "=", "≥", "≤", or "depends" to replace the "?". Choose the most specific relation that always holds and briefly explain why. Assume all probabilities are non-zero.

(i) $P(H = h, D = d) \; ? \; P(H = h)$

(ii) $P(H = h|D = d) \; ? \; P(H = h)$

(iii) $P(H = h|D = d) \; ? \; P(D = d|H = h)P(H = h)$

5) **Positive (semi-)definite matrices (20 pts).** Let $\mathbf{A}$ be a real, symmetric $d \times d$ matrix. We say $\mathbf{A}$ is *positive semi-definite* (PSD) if, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$. We say $\mathbf{A}$ is *positive definite* (PD) if, for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$. We write $\mathbf{A} \succeq 0$ when $\mathbf{A}$ is PSD, and $\mathbf{A} \succ 0$ when A is PD.

The *spectral theorem* (which we will assume without proof) says that every real symmetric matrix $\mathbf{A}$ can be expressed via the *spectral decomposition*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

where $\mathbf{U}$ is a $d \times d$ matrix such that $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ (called an orthogonal matrix), and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$. Multiplying on the right by $\mathbf{U}$ we see that $\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$. If we let $\mathbf{u}_i$ denote the i-th column of $\mathbf{U}$, we have $\mathbf{A}\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for each i. This expression reveals that the λ_i are eigenvalues of $\mathbf{A}$, and the corresponding columns are eigenvectors associated to λ_i. The eigenvalues constitute the "spectrum" of $\mathbf{A}$, and the spectral decomposition is also called the eigenvalue decomposition of $\mathbf{A}$.

Using the spectral decomposition, show that

(a) (10 pts) $\mathbf{A}$ is PSD iff $\lambda_i \geq 0$ for each i.

(b) (10 pts) $\mathbf{A}$ is PD iff $\lambda_i > 0$ for each i.

Hint: Use the identity

$$\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T,$$

which can be verified just by showing that the matrices representing the left and right hand sides have the same entries.

6) **Optimization (15 pts).** Recall that a function f is *convex* if the domain of f is convex and

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}),$$

for all $t \in [0,1]$ and $\mathbf{x}$, $\mathbf{y}$ in the domain of f. A function f is called strictly convex if the above relation is a strict inequality for all $t \in (0,1)$ and $\mathbf{x} \neq \mathbf{y}$ in the domain of f. A function $f(\mathbf{x})$ is concave iff $-f(\mathbf{x})$ is convex. A point $\mathbf{x}^*$ is a local minimizer of f if there exists an $\epsilon > 0$ such that $f(\mathbf{x}^* + \mathbf{y}) \geq f(\mathbf{x}^*)$ when $\mathbf{y}$ is such that $\|\mathbf{x}^* - \mathbf{y}\| \leq \epsilon$.

(a) (5 pts) Use the Hessian to establish that the sum of two convex functions is convex. You may assume f is twice continuously differentiable.

(b) (5 pts) Consider the function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where $\mathbf{A}$ is a symmetric $d \times d$ matrix. Derive the Hessian of f. Under what conditions on $\mathbf{A}$ is f convex? Strictly convex?

(c) (5 pts) Show that if f is twice-continuously differentiable and $\mathbf{x}^*$ is a local minimizer, then $\nabla^2 f(\mathbf{x}^*) \succeq 0$, i.e., the Hessian of f is positive semi-definite at the local minimizer $\mathbf{x}^*$ [2].

[2] A twice continuously differentiable function admits the quadratic expansion

$$f(\mathbf{x}) = f(\mathbf{y}) + [\nabla f(\mathbf{y})]^T(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\nabla^2 f(\mathbf{y})(\mathbf{x}-\mathbf{y}) + o\left(\|\mathbf{x}-\mathbf{y}\|^2\right),$$

where $o(t)$ denotes a function satisfying $\lim_{t\to 0}\frac{o(t)}{t} = 0$.