Homework #0

CSE 546: Machine Learning
Michael Ross
Due: 10/4/18 11:59 PM

1 Analysis

1. *[1 points]* A set $A \subseteq \mathbb{R}^n$ is *convex* if $\lambda x + (1-\lambda)y \in A$ for all $x, y \in A$ and $\lambda \in [0,1]$. A *norm* $\|\cdot\|$ over $\mathbb{R}^n$ is defined by the properties: i) non-negative: $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$, ii) absolute scalability: $\|a\,x\| = |a|\,\|x\|$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$, iii) triangle inequality: $\|x+y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

   a. Using just the definitions above, show that the set $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$ is convex for any norm $\|\cdot\|$.
   Answer:
   Let $A = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ and $x, y \in A$

   $\|\lambda x + (1-\lambda)y\| = \|\lambda x\| + \|(1-\lambda)y\| = |\lambda|\,\|x\| + |1-\lambda|\,\|y\|$
   $|\lambda| \in [0,1]$ by norm definition
   $\|x\| \in [0,1]$ by norm and set definitions
   $\|y\| \in [0,1]$ by norm and set definitions
   $\implies \|\lambda x + (1-\lambda)y\| \leq 1$

   $\lambda x + (1-\lambda)y$ is a linear combination of elements of $\mathbb{R}^n$
   $\implies \lambda x + (1-\lambda)y \in \mathbb{R}^n$

   Thus $\lambda x + (1-\lambda)y \in A$ for any x, y

   b. Show that $\left(\sum_{i=1}^n |x_i|^{1/2}\right)^2$ is or is not a norm.
   Answer:
   i) $x_i \in \mathbb{R}$
   $\implies |x_i|^{1/2} \in \mathbb{R}$
   $\implies \sum_{i=1}^n |x_i|^{1/2} \in \mathbb{R}$
   $\implies \left(\sum_{i=1}^n |x_i|^{1/2}\right)^2 \geq 0$

   ii) $\|ax\| = \left(\sum_{i=1}^n |ax_i|^{1/2}\right)^2$
   $= \left(\sum_{i=1}^n (|a||x_i|)^{1/2}\right)^2$
   $= \left(\sum_{i=1}^n |a|^{1/2}|x_i|^{1/2}\right)^2$
   $= \left(|a|^{1/2}\sum_{i=1}^n |x_i|^{1/2}\right)^2$
   $= |a|(\sum_{i=1}^n |x_i|^{1/2})^2$
   $= |a|\,\|x\|$

   iii)

2. *[1 points]* For any $x \in \mathbb{R}^n$, define the following norms: $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$, $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$. Show that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

3. *[1 points]* For possibly non-symmetric $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}$, let $f(x,y) = x^T\boldsymbol{A}x + y^T\boldsymbol{B}x + c$. Define $\nabla_z f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial z_1} & \frac{\partial f(x,y)}{\partial z_2} & \ldots & \frac{\partial f(x,y)}{\partial z_n} \end{bmatrix}^T$. What is $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$?

4. *[1 points]* Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $\mathbb{R}^{n\times n}$ symmetric matrices. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ have the exact same set of eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n$ with the corresponding eigenvalues $\alpha_1, \alpha_2, \cdots, \alpha_n$ for $\boldsymbol{A}$, and $\beta_1, \beta_2, \cdots, \beta_n$ for $\boldsymbol{B}$. Please write down the eigenvectors and their corresponding eigenvalues for the following matrices:

   a. $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$
   Answer:
   Eigenvectors: $\mathbf{u_1}, \mathbf{u_2}, ...\mathbf{u_n}$ Eigenvalues: $\alpha_1 + \beta_1, \alpha_2 + \beta_2, ...\alpha_n + \beta_n$

   b. $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{B}$
   Answer:
   Eigenvectors: $\mathbf{u_1}, \mathbf{u_2}, ...\mathbf{u_n}$ Eigenvalues: $\alpha_1 - \beta_1, \alpha_2 - \beta_2, ...\alpha_n - \beta_n$

   c. $\boldsymbol{E} = \boldsymbol{A}\boldsymbol{B}$
   Answer:
   Eigenvectors: $\mathbf{u_1}, \mathbf{u_2}, ...\mathbf{u_n}$ Eigenvalues: $\alpha_1\beta_1, \alpha_2\beta_2, ...\alpha_n\beta_n$

   d. $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$ (assume $\boldsymbol{A}$ is invertible)
   Answer:
   Eigenvectors: $\mathbf{u_1}, \mathbf{u_2}, ...\mathbf{u_n}$ Eigenvalues: $\beta_1/\alpha_1, \beta_2/\alpha_2, ...\beta_n/\alpha_n$

5. *[1 points]* A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive-semidefinite (PSD)* if $x^T\boldsymbol{A}x \geq 0$ for all $x \in \mathbb{R}^n$.

   a. For any $y \in \mathbb{R}^n$, show that yy^T is PSD.
   Answer:
   $x^T yy^T x = (x \cdot y)(y \cdot x) = (x \cdot y)^2 \geq 0$

   b. Let X be a random vector in $\mathbb{R}^n$ with covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}]$. Show that $\boldsymbol{\Sigma}$ is PSD.
   Answer:
   $x^T\Sigma x = x^T\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}]x$
   $= \mathbb{E}[x^T(X - \mathbb{E}[X])(X - \mathbb{E}[X])^{\mathrm{T}}x]$ since x is not random $x\mathbb{E}[X] = \mathbb{E}[xX]$
   Let $\gamma = (X - \mathbb{E}[X])^T x = x^T(X - \mathbb{E}[X])$
   $x^T\Sigma x = \mathbb{E}[\gamma^2]$
   $\gamma \in \mathbb{R} \implies \gamma^2 \geq 0 \implies \mathbb{E}[\gamma^2] \geq 0$
   $\implies x^T\Sigma x \geq 0$

   c. Assume $\boldsymbol{A}$ is a symmetric matrix so that $\boldsymbol{A} = \boldsymbol{U}\mathrm{diag}(\alpha)\boldsymbol{U}^T$ where $\mathrm{diag}(\alpha)$ is an all zeros matrix with the entries of α on the diagonal and $\boldsymbol{U}^T\boldsymbol{U} = I$. Show that $\boldsymbol{A}$ is PSD if and only if $\min_i \alpha_i \geq 0$. (Hint: compute $x^T\boldsymbol{A}x$ and consider values of x proportional to the columns of $\boldsymbol{U}$, i.e., the orthonormal eigenvectors).
   Answer:
   $x^TAx = x^T\boldsymbol{U}diag(\alpha)\boldsymbol{U}^Tx$
   Let $x = \boldsymbol{U}\beta$ where $\beta \in \mathbb{R}^n$
   $x^TAx = \beta^T\boldsymbol{U}^T\boldsymbol{U}diag(\alpha)\boldsymbol{U}^T\boldsymbol{U}\beta$
   $= \beta^T diag(\alpha)\beta$
   Switching to index notation
   $= \beta_i diag(\alpha)_{ij}\beta_j$
   $diag(\alpha)_{ij} = 0$ for $i \neq j$
   $\implies x^TAx = \sum_{i=1}^n \beta_i\beta_i\alpha_i$

6. *[1 points]* Let X and Y be real independent random variables with PDFs given by f and g, respectively. Let h be the PDF of the random variable $Z = X + Y$.

   a. Derive a general expression for h in terms of f and g
   Answer:
   $\mathbb{P}(Z = z) = \sum_{x+y=z} \mathbb{P}(X = x)\mathbb{P}(Y = y)$
   Generalize to PDFS
   $h(z) = \int_{x+y=z} f(x)g(y)dxdy$
   $h(z) = \int_{-\infty}^{\infty} f(x)g(z - x)dx$

b. If X and Y are both independent and uniformly distributed on $[0, 1]$ (i.e. $f(x) = g(x) = 1$ for $x \in [0, 1]$ and 0 otherwise) what is h, the PDF of $Z = X + Y$?

c. For these given explicit distributions, what is $\mathbb{P}(X \leq 1/2 | X + Y \geq 5/4)$?

7. *[1 points]* A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is Gaussian distributed with mean μ and variance σ^2. Given that for any $a, b \in \mathbb{R}$, we have that $Y = aX + b$ is also Gaussian, find a, b such that $Y \sim \mathcal{N}(0, 1)$.

Answer:
$\mathbb{E}[Y] = 0$
$a\mathbb{E}[X] + b = 0$
$a\mu + b = 0$
$\mu = -b/a$

$\mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = 1$
Since $\mathbb{E}[Y] = 0$
$\mathbb{E}[Y^2] = 1$
$\mathbb{E}[a^2X^2 + 2abX + b^2] = 1$
$a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 = 1$
$a^2\mathbb{E}[X^2] - a^2\mathbb{E}[X]^2 + a^2\mathbb{E}[X]^2 + 2ab\mathbb{E}[X] + b^2 = 1$
$a^2\sigma^2 + a^2\mu^2 + 2ab\mu + b^2 = 1$
Plug $\mu = -b/a$ in
$a^2\sigma^2 + b^2 - 2b^2 + b^2 = 1$
$\implies a = 1/\sigma$
$\implies b = -\mu/\sigma$

8. *[1 points]* If $f(x)$ is a PDF, we define the cumulative distribution function (CDF) as $F(x) = \int_{-\infty}^{x} f(y)dy$. For any function $g : \mathbb{R} \to \mathbb{R}$ and random variable X with PDF $f(x)$, define the expected value of $g(X)$ as $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(y)f(y)dy$. For a boolean event A, define $\mathbf{1}\{A\}$ as 1 if A is true, and 0 otherwise. Thus, $\mathbf{1}\{x \leq a\}$ is 1 whenever $x \leq a$ and 0 whenever $x > a$. Note that $F(x) = \mathbb{E}[\mathbf{1}\{X \leq x\}]$. Let $X_1, \ldots, X_n$ be *independent and identically distributed* random variables with CDF $F(x)$. Define $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$.

a. For any x, what is $\mathbb{E}[\widehat{F}_n(x)]$?
Answer:
$\mathbb{E}[\widehat{F}_n(x)] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}]$
$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{1}\{X_i \leq x\}]$
$= \frac{1}{n}\sum_{i=1}^{n} F(x)$
$= F(x)$

b. For any x, show that $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = \frac{F(x)(1-F(x))}{n}$
Answer:
$\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = \mathbb{E}[\widehat{F}_n(x)^2 - 2\widehat{F}_n(x)F(x) + F(x)^2]$
$= \mathbb{E}[\widehat{F}_n(x)^2] - 2\mathbb{E}[\widehat{F}_n(x)]F(x) + F(x)^2]$

$\widehat{F}_n(x)^2 = \frac{1}{n^2}(\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\})^2$
$= \frac{1}{n^2}(\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}\mathbf{1}\{X_1 \leq x\} + ... + \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}\mathbf{1}\{X_n \leq x\})$
$= \frac{1}{n^2}(\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} + ... + \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\})$
$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$
$= \widehat{F}_n(x)$

$\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = F(x) - 2F(x)F(x) + F(x)^2$
$= F(x)(1 - F(x))$

c. Using part b., show that $\sup_{x \in \mathbb{R}} \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$.
Answer:

$$\sup_{x\in\mathbb{R}} \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = \sup_{x\in\mathbb{R}} \frac{1}{n}(F(x) - F(x)^2)$$

$$\partial_{F(x)}(\tfrac{1}{n}(F(x) - F(x)^2)) = \tfrac{1}{n}(1 - 2F(x))$$
$$\arg\max_{F(x)}(\tfrac{1}{n}(F(x) - F(x)^2)) = 1/2$$
$$\max\ (\tfrac{1}{n}(F(x) - F(x)^2)) = \tfrac{1}{4n}$$
$$\implies \sup_{x\in\mathbb{R}} \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$$

2 Programming

9. *[2 points]* Two random variables X and Y have equal distributions if their CDFs, F_X and F_Y, respectively, are equal: $\sup_x |F_X(x) - F_Y(x)| = 0$. The central limit theorem says that the sum of k independent, zero-mean, variance-$1/k$ random variables converges to a Gaussian distribution as k goes off to infinity. We will study this phenomenon empirically (you will use the Python packages Numpy and Matplotlib). Define $Y^{(k)} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} B_i$ where each B_i is equal to -1 and 1 with equal probability. It is easy to verify (you should) that $\frac{1}{\sqrt{k}} B_i$ is zero-mean and has variance $1/k$.

   a. For $i = 1, \ldots, n$ let $Z_i \sim \mathcal{N}(0, 1)$. If $F(x)$ is the true CDF from which each Z_i is drawn (i.e., Gaussian) and $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Z_i \leq x\}$, use the homework problem above to choose n large enough such that $\sup_x \sqrt{\mathbb{E}[(\widehat{F}_n(x) - F(x))^2]} \leq 0.0025$, and plot $\widehat{F}_n(x)$ from -3 to 3. (Hint: use `Z=numpy.random.randn(n)` to generate the random variables, and `import matplotlib.pyplot as plt`; `plt.step(sorted(Z), np.arange(1,n+1)/float(n))` to plot).
   Answer:

```
import matplotlib.pyplot as plt
import numpy as np

supMax=0.0025
n=int(1.0/(2.0*supMax)**2)
Z=np.random.randn(n)

plt.step(sorted(Z),np.arange(1,n+1)/float(n))
plt.xlabel("Observations")
plt.ylabel("Probability")
plt.xlim((-3,3))
plt.ylim((0,1))
plt.grid()
plt.show()
```

   b. For each $k \in \{1, 8, 64, 512\}$ generate n independent copies $Y^{(k)}$ and plot their empirical CDF on the same plot as part a. (Hint: you can use `np.sum(np.sign(np.random.randn(n, k))*np.sqrt(1./k), axis=1)` to generate n of the $Y^{(k)}$ random variables.)
   Answer:

```
import matplotlib.pyplot as plt
import numpy as np

supMax=0.0025
        n=int(1.0/(2.0*supMax)**2)
Z=np.random.randn(n)

k=[1,8,64,512]
i=1000
```

```
    for j in range(len(k)):
        x=np.sum(np.sign(np.random.randn(i, k[j]))*np.sqrt(1./k[j]), axis=1)
        plt.step(sorted(x),np.arange(1,i+1)/float(i))

    plt.step(sorted(Z),np.arange(1,n+1)/float(n))
    plt.xlabel("Observations")
    plt.ylabel("Probability")
    plt.xlim((-3,3))
    plt.ylim((0,1))
    k.append('Gaussian')
    plt.legend(k)
    plt.grid()
    plt.show()
```

Be sure to always label your axes. Your plot should look something like the following (Tip: checkout `seaborn` for instantly better looking plots.)