Midterm: CAP6610 Fall '20
Time: 120 minutes
Open Book; Open Notes

1. (10 pts) {Mathematical Probability Theory} Let X be a random variable with distribution function:

$$F_X(x) = 0 \text{ if } \quad -\infty < x \leq 3 \tag{1}$$

$$F_X(x) = \frac{1}{2}(x-3) \text{ if } \; 3 < x \leq 5 \tag{2}$$

$$F_X(x) = 1 \text{ if } \; 5 < x < \infty \tag{3}$$

Let Y be a random variable with distribution function:

$$F_Y(y) = 0 \text{ if } \quad -\infty < y \leq 0 \tag{4}$$

$$F_Y(y) = y \text{ if } \; 0 < y \leq 1 \tag{5}$$

$$F_Y(y) = 1 \text{ if } \; 1 < y < \infty \tag{6}$$

(i) Solve for the density functions $f_X(x)$ and $f_Y(y)$ of random variables X and Y, respectively. Draw/sketch them.

(ii) Let X and Y be *independent* random variables. Recall that mathematically speaking this corresponds to $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all x, y. Solve for the density function $f_{X,Y}(x,y)$ and draw/sketch it.

2. (10 pts) {Mathematical Probability Theory} Continue with the assumptions of Question 1. Let random variable $Z = X + Y$. Solve for the density functions $f_Z(z)$. Draw/sketch it.

3. (10 pts) {Mathematical Probability Theory} Continue with the assumptions of Questions 1 and 2. Solve for the *conditional density function* $f_{Z|Y}(z|y)$. Draw/sketch it.

4. (10 pts) {Sigma Algebra} Recall that a sigma algebra $\mathcal{F}$ on a sample space Ω contains the empty set ϕ and is closed under countable union and complementation. We are going to build the smallest sigma algebra on $\mathbb{R}$ that contains all sets $(-\infty, a]$ for any $a \in \mathbb{R}$; which as you will recall is also known as the sigma algebra *generated* by the aforementioned sets.

Prove that the set $(3.5, 4.5) - \{4.0\}$ is in the sigma algebra. Stated in words, this set includes all real numbers between 3.5 and 4.5 but not including 3.5 and 4.5, and in addition does not include the point 4.0.

Your proof will look like if set A is in the sigma algebra then set B has to be in the sigma algebra, and if set B is in the sigma algebra then set C has to be in the sigma algebra and so on.

5. (10 pts) {Risk functional approach to regression} Recall that in class, we formulated $R(\alpha) = \int (y - f(x;\alpha))^2 dF_{XY}$ as the risk functional to be minimized for regression, following which we turned it into the empirical risk functional $R_{emp}(\alpha) = \Sigma_{i=1}^{N}(y_i - f(x_i;\alpha))^2/N$ when only i.i.d samples are available. Here α parameterizes the concept class/hypothesis space

of regression functions to choose from and (x_i, y_i) for $i = 1 \ldots N$ are the N independent/dependent variable sample tuples.

Assume that our concept class is the set of all possible functions of the form $y = ax + b$. Solve for a and b for the dataset: $(x_1, y_1) = (-1, 1)$; $(x_2, y_2) = (0, 0)$; $(x_3, y_3) = (1, 2)$. Note that the number of samples $N = 3$ here.

6. (10 pts) {Risk functional approach to regression} Continue with the assumptions of Question 5. Assume this time that our concept class is the set of all possible functions of the form $y = ax^2 + bx + c$. Solve for a, b and c for the same dataset.

7. (10 pts) {Risk functional approach to density estimation} Recall that in class, we formulated $R(\alpha) = -\int \ln p(x; \alpha) dF_X$ as the risk functional to be minimized for density estimation, following which we turned it into the empirical risk functional $R_{emp}(\alpha) = -\Sigma_{i=1}^{N} \ln p(x_i; \alpha)/N$ when only i.i.d samples are available. Here α parameterizes the concept class/hypothesis space of densities to choose from and x_i for $i = 1 \ldots N$ are the N samples.

   Assume that our concept class is the set of all possible *uniform density functions* over the range $[a, b]$. That is, the density: $f(x; a, b) = 0$ if $x < a$, $f(x; a, b) = 1/(b - a)$ if $a \leq x \leq b$, and $f(x; a, b) = 0$ if $x > b$. Note that in this case the parameter α is comprised of the pair of real numbers (a, b).

   Given samples $x_1, \ldots x_N$, where $x_i \in \mathbb{R}$, the risk functional approach then reduces to minimizing $R_{emp}(a, b) = -\Sigma_{i=1}^{N} \ln f(x_i; a, b)/N$ with respect to a, b.

   Given samples $x_1, \ldots x_N$ what are the values of a and b? Explain your answer.

   Hint: Draw pictures to get a sense of the problem first.

8. (10 pts) {Multi Layer Perceptron} Recall that a single perceptron computes the function $sgn(\sum_{i=1}^{n} w_i x_i + b)$ where $x_1 \ldots x_n$ are the inputs, b is the bias, and sgn is the sign function that outputs $+1$ or -1 depending on whether its input is greater or less than 0.

   Your goal in this problem is to hand craft a 2-layer perceptron network that for boolean inputs x_1, x_2, x_3, outputs $f(x_1, x_2, x_3) = (x_1 \wedge \neg x_2 \wedge x_3) \vee (\neg x_1 \wedge \neg x_3)$. Each boolean input is set to $+1$ if it is true and -1 if it is false.

   $\neg$ stands for *not*, $\wedge$ stands for *and*, and $\vee$ stands for *or*. $A \wedge B$ is true only if both A is true and B is true, and is false otherwise. $A \vee B$ is true if either A is true or B is true, and is false otherwise. $\neg A$ is true if A is false and vice versa.

   Argue why the weights and biases you have chosen in your 2-layer perceptron network give the correct overall answer.

9. (10 pts) {Artificial Neural Network} In class, we replaced the sgn function in the perceptron with the logistic function $f(x) = \frac{1}{1+e^{-x}}$ in the artificial neural network to devise a *gradient descent* update. However, whereas the sgn function takes a value -1 or +1, the logistic function takes values ranging from 0 to 1. Derive the gradient descent update rule for the perceptron when the logistic function is scaled and shifted, so that it now ranges from -1 to +1 (as noted in class, this is related to $\tanh$, the tan hyperbolic function). In particular, make sure that $\frac{\partial o}{\partial net}$, as derived in class, is represented as a function of o.

10. (10 pts) $\{$Artificial Neural Network$\}$ Deep networks, nowadays, routinely introduce $\max$ nodes into their network. Given N inputs $x_1, \ldots, x_N$ and corresponding weights $w_1, \ldots, w_N$ the output o of a max node is $\max_{i=1 \ldots N}\{w_1 x_1, w_2 x_2, \ldots, w_N x_N\}$. In words, the output o chooses (and is therefore equal to) the maximum over all the weighted inputs.

    What is the value of $\partial o / \partial w_i$. Explain your answer.