ENM531 Midterm exam (Spring 2020)

Instructor: Paris Perdikaris

Thursday, April 2, 2020

Question 1 (5 pts)

Give an example of a case where you would like to use an L^1 penalty rather than an L^2 penalty in the loss function of a supervised model. Please provide some explanation on it.

Question 2 (10 pts)

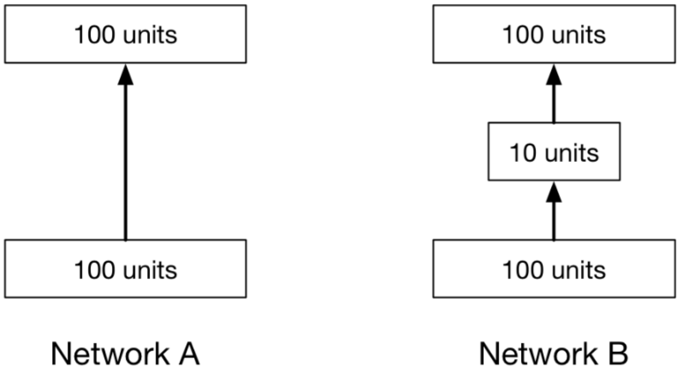
Consider a pair of discrete random variables X and Y whose joint distribution is as follows:

	$Y = 0$	$Y = 1$
$X = 0$	0	0.5
$X = 1$	0.25	0.25

(a) Compute the joint entropy $\mathcal{H}(X, Y)$ (5 pts).
(b) Compute the conditional entropy $\mathcal{H}(Y|X)$ (5 pts).

Question 3 (10 pts)

Consider the following two multilayer perceptrons, where all of the layers use linear activation functions.



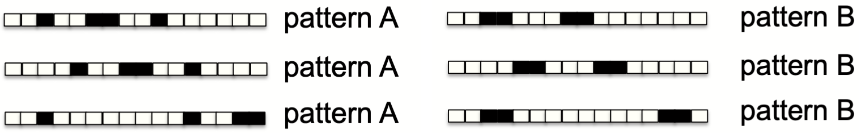
(a) Give one advantage of network A over network B (5 pts).
(b) Give one advantage of network B over network A (5 pts).

Question 4 (15 pts)

Consider the following convolutional layer in a conv net. The input has a spatial dimension 12×12, and has 10 channels. The convolution kernels are 3×3, the stride is 2, and we use “valid" convolution, which means that each output neuron only looks at image regions that lie entirely within the spatial bounds of the input. The output dimension is 5×5, with 20 channels.
For this question, you don't need to show your work or justify your answer, but doing so may help you get partial credit.
(a) How many weights are required for this convolution layer? (10 pts)
(b) Suppose we instead make this a locally connected layer, i.e. don't use weight sharing. How many weights are required? (5 pts)

Question 5 (15 pts)

Design your own convolutional neural network. This task is to distinguish the following two patterns collected on 16 spatial positions. The inputs are 16-dimensional binary vectors, where black indicates 1 and white indicates 0.
You are asked to design your own convolutional neural network architecture to classify these patterns.



- First we have a convolution layer with a single convolution kernel of size 3 with weights $w = [w_1, w_2, w_3]^T$ and bias b. Unlike in ordinary conbolution layers, this one will use wrap-around (consider the whole sequence as a circle). The output of this layer has 16 units.
- We apply the ReLU activation function to this layer.
- We pool together the activations by taking the sum. Call this value z.
- We threshold the result at a value r. If $r \leq z$, it is classified as B, otherwise it's classified as A.

Your task is to choose the weights w, bias b, and threshold r to correctly separate all instances of the patterns. You are not required to show your work, but explaining your reasoning may help you get partial credit.

Question 6 (20 pts)

Consider the problem of MAP estimation for the mean μ of a Gaussian distribution with known standard deviation σ. For the prior distribution, we will use a Gaussian distribution with mean 0 and standard deviation γ.

(a) Determine the function that we need to maximize. You do not need to determine the constant terms explicitly (10 pts).
(b) Determine the optimal value of μ by setting the derivative to 0. (You do not need to justify why it is a maximum rather than a minimum.) (10 pts)

Question 7 (25 pts + 10 bonus pts)

Posterior distribution tries to balance prior information and data-fit: let y be the number of heads in n tosses of a coin, whose probability of heads is θ.
(a) If your prior distribution for θ is uniform on the range $[0, 1]$, derive your prior predictive distribution for y,

$$p(y = k) = \int_0^1 p(y = k|\theta)p(\theta)d\theta$$

for each $k = 0, 1, ..., n$. (10 pts)
(b) Suppose you assign a Beta(α, β) prior distribution for θ, and then you observe y heads out of n tosses. Show algebraically that your posterior mean of θ always lies between your prior mean, $\frac{\alpha}{\alpha+\beta}$, and the observed relative frequency of heads, namely $\frac{y}{n}$. (15 pts)
(c) Show that, if the prior distribution on θ is uniform, the posterior variance of θ is always less than the prior variance. (Bonus 10 pts)

Hint: Some useful formulas for the Gamma and Beta functions:

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t)dt, \tag{1}$$

$$\Gamma(x + 1) = x! \quad \text{when} \quad x \in N^+, \tag{2}$$

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt, \tag{3}$$

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \tag{4}$$

Hint: Some useful distributions:

$$\text{Beta distribution pdf Beta(x; a, b)} = \frac{x^{a-1}(1 - x)^{b-1}}{B(a, b)} \tag{5}$$

$$\text{Gamma distribution pdf Gamma(x; a, b)} = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \tag{6}$$