ENM 531 Midterm 2020

Time: 3 hrs.
Please submit your answers in the format of PDF.
No snapshot formats are accepted.

For all questions, if your answer is close to the solution or your explanation makes sense, you will get portion of the credits.

1 Question 1 (5 pts)

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

L^1 regularization encourages sparse weight vectors, while L^2 encourages all the weights to be small but nonzero. We'd prefer L^2 in situations where all the features are likely to be relevant. Example could include polynomian regression, classifying MNIST digits from pixels, etc. We'd like to use L^1 in situations where we want to encourage sparsity, have less features, etc.

2 Question 2 (10 pts)

Consider a pair of 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).

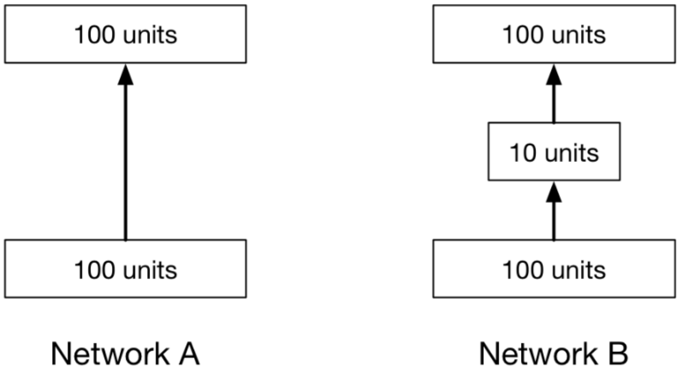
log function with base 2 is accepted if you mention that in your answer. That way, the answer for (a) should be 1.5 and (b) is 0.5. However, if you do not mention it, we take 1 credit for each question.

(a) The joint entropy of multiple categorical random variables is the same as the entropy of a single categorical random variable with the same set of probabilities. So, in this case, the entropy is: $-0.5\log(0.5) - 0.25\log(0.25) - 0.25\log(0.25) = 1.0397$

(b) $\mathcal{H}(Y|X) = -\sum p(x,y)\log\frac{p(x,y)}{p(x)}$
$= \sum p(x)H(Y|X=x)$
$= 0 - 0.5\log(0.5/0.5) - 0.25\log(0.25/0.5) - 0.25\log(0.25/0.5)$
$= -0.5\log(0.5)$
$= 0.346$

3 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).

(a) A has more connections than B, so it will be able to preserve more information.
(b) B has fewer connections, so it's less prone to overfitting. B has fewer connections, so backprop requires fewer operations. B has a bottleneck layer, so the network is forced to learn a compact representation (like an autoencoder).

If you say B is having more parameters than A, this is definitely wrong.

4 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

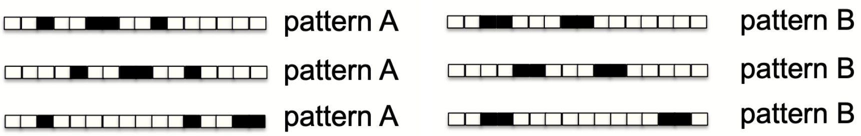
may help you get partial credit.
(a) How many weights are required for this convolution layer? (10 pts)
(b) Suppose we instead make this a locally connected layer, i.e. don't use weight sharing. How many weights are required? (5 pts)

(a) A convolutional layer with N input channels, M output channels, and $K \times K$ spatial extent requires MNK^2 weights. Hence, we need: $10 \times 20 \times 3 \times 3 = 1800$ weights.

(b) The locally connected layer has the same pattern of connections as the convolution layer but each of the $5 \times 5 = 25$ output locations will have its own separate set of weights. Hence, the total number of weights is $25 \times 1800 = 45000$.

5 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 $\boldsymbol{w} = [w_1, w_2, w_3]^T$ 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 $\boldsymbol{w}$, 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.

There are lots of possible solutions. One is: $\boldsymbol{w} = [1, 1, 0]^T$, $b = -1$ and $r = 1.5$. The convolution layer will have an output of 0 elsewhere. Hence, it will activate in 2 locations for pattern B and 1 location for pattern A. If we threshold it at 1.5, it will separate instances of A and B.

For this question, as long as your answer and explanation make sense, you can get the credit.

6 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)

(a)

$$\begin{aligned}\log p(\mu|\mathcal{D}) &= const + \log p(\mu) + \log p(\mathcal{D}|\mu) &(1)\\ &= const + \log \mathcal{N}(\mu; 0, \gamma) + \sum_{i=1}^{N} \log \mathcal{N}(x^{(i)}; \mu, \sigma) &(2)\\ &= const - \frac{1}{2\gamma^2}\mu^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^{(i)} - \mu)^2 &(3)\end{aligned}$$

(b)

$$\begin{aligned}\frac{dJ}{d\mu} &= -\frac{\mu}{\gamma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N}(x^{(i)} - \mu) &(4)\\ &= -(\frac{1}{\gamma^2} + \frac{N}{\sigma^2})\mu + \frac{1}{\sigma^2}\sum_{i=1}^{N} x^{(i)} &(5)\end{aligned}$$

Letting this to 0, we get:

$$\mu = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{N} x^{(i)}}{\frac{1}{\gamma^2} + \frac{N}{\sigma^2}} \qquad (6)$$

7 Question 7 (25 pts + 10 extra 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:

$$\Gamma(x)=\int_0^\infty t^{x-1}\exp(-t)dt, \tag{7}$$

$$\Gamma(x+1)=x! \quad \text{when} \quad x\in N^+, \tag{8}$$

$$\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt, \tag{9}$$

$$\mathrm{B}(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{10}$$

Hint: Some useful distributions:

$$\text{Beta distribution pdf Beta(x; a, b)=}\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \tag{11}$$

$$\text{Gamma distribution pdf Gamma(x; a, b)=}\frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx) \tag{12}$$

(a)

$$\begin{aligned}
p(y=k) &= \int p(y=k|\theta)d\theta && (13)\\
&= \int \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta && (14)\\
&= \binom{n}{k}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} && (15)\\
&= \frac{\Gamma(n+1)\Gamma(k+1)}{\Gamma(n-k+1)}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} && (16)\\
&= \frac{\Gamma(n+1)}{\Gamma(n+2)} && (17)\\
&= \frac{1}{n+1} && (18)
\end{aligned}$$

(b) As we have the posterior distribution of θ as:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} && (19)\\ p(\theta|y) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{y}(1-\theta)^{n-y} && (20)\\ &\propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1} && (21)\\ &\sim \mathrm{Beta}(y+\alpha,\beta+n-y) && (22)\end{aligned}$$

In other words, mean of the posterior is $\frac{\alpha+y}{\alpha+\beta+n}$.
It could be easily realized that all the possible values between the prior mean and the frequency inference could be parameterized as:

$$\gamma = b\frac{y}{n} + (1-b)\frac{\alpha}{\alpha+\beta} \tag{23}$$

where $b \in [0,1]$. Letting that $\gamma = \frac{\alpha+y}{\alpha+\beta+n}$ and solve the equation to find $\hat{b}$. We would expect that $\hat{b} \in [0,1]$. In this case:

$$\begin{aligned} \frac{\alpha+y}{\alpha+\beta+n} &= \hat{b}\frac{y}{n} + (1-\hat{b})\frac{\alpha}{\alpha+\beta} && (24)\\ n(\alpha+y)(\alpha+\beta) &= \hat{b}[(\alpha+\beta+n)(\alpha+\beta)y-(\alpha+\beta+n)n\alpha]+(\alpha+\beta+n)n\alpha && (25)\\ n(\alpha y+\beta y-n\alpha) &= \hat{b}(\alpha+\beta+n)(\alpha y+\beta y-n\alpha) && (26)\\ \hat{b} &= \frac{n}{\alpha+\beta+n} && (27)\end{aligned}$$

And truly $\hat{b} \in [0,1]$ for $\alpha > 0, \beta > 0, n > 0$. In this case, we can conclude that posterior mean of θ always lies between your prior mean $\frac{\alpha}{\alpha+\beta}$ and the observed relative frequency of heads $\frac{y}{n}$.

(c) For the uniform distribution we have $\alpha = 1$ and $\beta = 1$.
In this case, we find that the posterior distribution of θ on y is also Beta distribution:

$$p(\theta|y) \sim \mathrm{Beta}(y+1, n-y+1) \tag{28}$$

In this case, we can find the variance be

$$\mathrm{var}(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}$$

with also the prior distribution of $p(\theta) = 1$. We then compute the variance of prior θ:

$$\mathrm{var}(\theta) = \int_0^1 (\theta-0.5)^2 d\theta = \frac{1}{3}(\theta-0.5)^3|_0^1 = \frac{1}{12} \tag{29}$$

we have the inequality that:

$$(y+1)(n-y+1) < \frac{(n-y+1+y+1)^2}{4} = \frac{(n+2)^2}{4}$$

So, we can summarize that:

$$\mathrm{var}(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)} < \frac{(n+2)^2}{4(n+2)^2(n+3)} = \frac{1}{4(n+3)} < \frac{1}{12} = \mathrm{var}(\theta) \tag{30}$$

8 Question 8

Beta-binomial distribution and Bayes' prior distribution: suppose y has a binomial distribution for given n and unknown parameter θ, where the prior distribution of θ is Beta(α, β).
(a) Find $p(y)$, the marginal distribution of y, for $y = 0, ..., n$ (unconditional on θ). This discrete distribution is known as the beta-binomial, for obvious reasons.
(b) Show that if the beta-binomial probability is constant in y, then the prior distribution has to have $\alpha = \beta = 1$.

(a) We can use the Bayesian rule to compute the marginal distribution of $p(y)$ as:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{31}$$

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)} \tag{32}$$

$$= \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{n}{y}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{\frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}} \tag{33}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(\alpha+\beta+n)} \tag{34}$$

Here we use the reality that the posterior distribution $p(\theta|y)$ is Beta distribution and use the form of it.

(b) We can observe from the above expression. The terms that involve y are:

$$\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)} \tag{35}$$

The rest terms are just constant. In order to make this above term equals to constant, we need them be independent of y. Due to the property of Gamma function, we would know that $\frac{\Gamma(y+\alpha)}{\Gamma(y+1)}$ and $\frac{\Gamma(n-y+\beta)}{\Gamma(n-y+1)}$ should be constant each. In this case, we would only be able to let $\alpha = 1$ and $\beta = 1$ for these terms to cancel out.