CS 337 Autumn 2019 | Endsem

Date/Time: August 30, 2019, 8 am to 9 am

Total marks: **15**

Name: _____

Roll number: _____

Instructions

- This is a closed notes quiz which should be completed individually.
- No form of collaboration or discussion is allowed.
- No laptops or cell phones are allowed.
- Write your name and roll number on the top of this page.
- This exam consists of 4 problems. The maximum possible score is 15.
- Write your answers legibly in the space provided on the exam sheet. (If necessary, use/ask for extra sheets to work out your solutions. These sheets will not be graded.)
- Work efficiently. The questions are not sorted in any order of difficulty. Try and attempt the easier ones first, so that you are not bogged down by the harder questions.
- Good luck!

Question	Score
Problem 1: Matching methodologies	/2
Problem 2: Maximum likelihood and Poisson	/2
Problem 3: Please help a shop owner with probabilistic modeling!	/6
Problem 4: Ridge Regression and Error Minimization	/5
Total	/50

Problem 1: Matching methodologies

For the following problems on the left hand side, pick a trick or methdology appropriate for it from the right hand side. No justification is needed. Your answer can be a list of the form (x) ⇒ (y) (2 marks)

Problem	Methodology
(1) Random variable X is defined in terms of random variable Y. Derive the pdf of X using the pdf of Y	**(a)** Using a monotonically non-increasing transformation
(2) Maximizing an objective function	**(b)** Constraining parameters
(3) Computing a joint distribution	**(c)** Using Moment Generating Function
(4) Avoiding overfitting	**(d)** Testing for independence

SOLUTION: Solution. **(1) ⇒ (c), (2) ⇒ (a), (3) ⇒ (d), (4) ⇒ (b)** For avoiding overfitting, we discussed how constraining the magnitude of each component of the weight vector can help avoid overfitting; overfitting manifests itself by weights increasing in magnitude). As for maximimizing an objective function, we discussed in Tutorial 1, problem 3 how the argmax is invariant to monotonically increasing transformations of the objective function. The other two are from basic probability theory. A joint distribution is product of marginals ($P(X,Y) = P(X)P(Y)$) if and only if X and Y are independent.

Problem 2: Maximum likelihood and Poisson

Let $x_1 = x_2 = x_3 = 1$, $x_4 = x_5 = x_6 = 2$ be a random sample from a Poisson random variable with mean θ ,where $\theta \in \{1, 2\}$. Recall that the probability mass function (pmf) of a Poisson random variable X with parameter θ is $\Pr(X = k) = e^{-\theta}\frac{\theta^k}{k!}$.

The maximum likelihood estimator of θ is equal to which of the following options? A brief justification will carry 1 mark.

(A) 1
(B) 1.5
(C) 1.58
(D) 2

(2 marks)

SOLUTION: Solution. (D): The solution seems to be 1.5 on first impression, but the actual answer is 2. The question has a constraint in the parameter choice, in the sense that not all values were allowed for the maxima search. The set specified for maximizing the likelihood was 1,2. Likelihood function

$$\mathcal{L}(\theta) = \frac{\theta^1 e^{-\theta}}{1!}.\frac{\lambda^1 e^{-\theta}}{1!}.\frac{\theta^1 e^{-\theta}}{1!}.\frac{\theta^2 e^{-\theta}}{2!}.\frac{\theta^2 e^{-\theta}}{2!}.\frac{\theta^2 e^{-\theta}}{2!} = \frac{\theta^9 e^{-6\theta}}{8}$$

. Verify that,

$$\mathcal{L}(1) \leq \mathcal{L}(2)$$

Gaussian, $\mu = 100$, $\sigma = 3$ (shop owner) , $\sigma = 1$

$P(D|\theta)P(\theta)$

$P(\mu|x_1, \ldots, x_5)$

.

Problem 3: Please help a shop owner with probabilistic modeling!

A shop owner believes that a sack of Jawar (a kind of grain) weighs 100 kgs and believes that his estimate could be off with roughly 68% probability within (plus/minus or $\pm$) 3 kgs. Assuming no other constraint on the probability distribution of the weight of the sack, what would you choose as the probability density function (pdf)? (1 marks)

Whereas, the whole sale dealer from whom the shop owner purchased his sack tells him that the weight could be off with roughly 68% probability within (plus/minus or $\pm$) 1 kg. Again, assuming no other constraint on this prior belief, what would you choose as the pdf for modeling this prior distribution? (1 marks)

He asks each of his 5 assistants to measure the weight of the sack, but since the shop only has a small weighing scale, their measurements are somewhat different: 102 kg, 99 kg, 103 kg, 97 kg and 99 kg respectively. How would you model the belief of the shopkeeper after having heard these observations? (2 marks)

As per your answer, to what *degree* will the shopkeeper now (after the observations) believe the weight of the sack to be 100 kgs? Your measure of *degree* can be probability based.

Is this degree greater than or less than the degree of his belief that the weight of sack was 100 kgs BEFORE he asked his 5 assistants to do their measurement? (1 marks)

Justify each step of your answer. Wherever possible without use of calculators etc, compute your answer.

Answer following parts very briefly: How would each step of your answer change if the weights were not measured in kgs but in pounds? And how would each step of your answer change if the weights were not measured in kgs but in handfuls? (1 marks)

Solution:

Recall that the normal distribution has the maximum entropy among all real-valued distributions with a specified variance.

It makes sense to model the weight of the sack as a normal distribution with an unknown mean μ and known standard deviation (squared) of 1 kg. That is, $\sigma^2 = 1$. The maximum likelihood estimate μ_{ML} for the mean μ is therefore the average of 102, 99, 103, 97 and 99, which is $\mu_{ML} = 100$ kg. (**1 Mark**)

By the 3-sigma rule, given no extra information, it makes most sense to model the prior distribution on μ as $\mathcal{N}(100, 9)$ (that is, $\mu_0 = 100$ and $\sigma_0^2 = 9$) (**1 Mark**)

Thus, the posterior distribution on the unknown mean will be (for $m = 5$)

$\Pr(\mu|x_1 \ldots x_5) = \mathcal{N}(\mu_5, \sigma_5^2)$ such that

$\mu_5 = (\frac{\sigma^2}{5\sigma_0^2 + \sigma^2}\mu_0) + (\frac{5\sigma_0^2}{5\sigma_0^2 + \sigma^2}\hat{\mu}_{ML})$ and $\frac{1}{\sigma_5^2} = \frac{1}{\sigma_0^2} + \frac{5}{\sigma^2}$

Since $\mu_0 = \mu_{ML} = 100$, $\mu_5 = 100$. Since $\frac{1}{\sigma_5^2} > \frac{1}{\sigma_0^2}$, $\sigma_5^2 < \sigma_0^2$, implying that $p(100|x_1 \ldots x_5) = \mathcal{N}(100, \sigma_5^2) > p(100|\emptyset) = \mathcal{N}(100, \sigma_0^2)$.

(**2 Marks**)

That is, the degree of belief AFTER the measurements of the 5 assistants will be greater than the degree of belief BEFORE the measurements of the 5 assistants. (**1 Mark**)

The entire analysis (and answers) will remain the same even if all weights are measured in pounds, since pounds is a continuous unit of measurement. For a discrete unit of measurement such as 'number of handfuls', the distribution would be a count distribution such a binomial and the prior could be chosen as a Beta. (**1 Mark**)

Problem 4: Ridge Regression and Error Minimization

Prove the following Claim: The sum of squares error on training data using the weights obtained after minimizing ridge regression objective *is greater than or equal to* the sum of squares error on training data using the weights obtained after minimizing the ordinary least squares (OLS) objective.

More specifically, if ϕ and $\mathbf{y}$ are defined on the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1)...(\mathbf{x}_m, y_m)\}$ as

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & ..... & \phi_n(\mathbf{x_1}) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & ..... & \phi_n(\mathbf{x}_m) \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \tag{2}$$

and if

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

and

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

then you should prove that

$$||\phi\mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi\mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

(4 marks)

If it is the case that ridge regression leads to greater error than ordinary least squares regression, then why should one be interested in ridge regression at all? (1 marks)

SOLUTION: If

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

then by definition of argmin,

$$||\phi\mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi\mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

Also, one can reformulate

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

as

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } ||\mathbf{w}||_2^2 \leq \theta$$

for some θ corresponding to a value of λ. The solution to a constrained minimization problem will always be greater than or equal to its unconstrained counterpart.

Ridge regression is still acceptable since ridge regression incorporates prior (as per Bayesian interpretation). The idea is ultimately to do well on unseen (test) data. Therefore, higher training error might be acceptable if test error can be lowered.