Midterm Exam

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2020

Monday February 10th

NAME: Yushang Lai

- Please write your name above. Don't turn to the next page until instructed to do so.
- **Closed book: no notes, no books, no electronic devices, etc.**
- 4 problems, each worth 20 points
- If you don't understand a question please read it carefully and at least twice. If you believe there is a typo or error in the statement of the question please raise your hand. Please do not raise your hand to try to get hints from the instructor—this is distracting and unfair to other students.
- There are some extra pages at the end if you need to continue your solution to a particular problem and need more space. Please indicate this clearly.
- You can use the backs of pages for scratch work - these won't be graded.
- If you need extra blank pages for your solutions please raise your hand to request some.

Problem 1: (20 points)

MAP MPE

1. Define precisely (with a formula) what we mean by the maximum a posteriori estimate for a parameter θ given a likelihood $P(D|\theta)$ and a prior $P(\theta)$.

by bayesian $p(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$ where $P(D)$ is constant

thus $p(\theta|D) \propto P(D|\theta)P(\theta)$

maximum a posterior estimate : MPE

we want $\underset{\theta}{\text{argmax}}\, E(P(\theta|D))$: find a θ s.t. $E(P(\theta|D))$ is maximum

this θ is the one mean

2. Consider a d-dimensional vector $\underline{x}$ where $p(\underline{x})$ is a multivariate Gaussian (Normal) density with parameters mean μ and covariance matrix Σ. State exactly how many parameters are in this model.

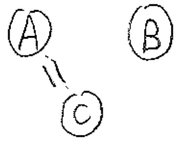
for mean μ, since d-dimensional. it has d parameters

for cov matrix Σ is d by d and symmetric thus need

$1+2+\cdots+d = \frac{(1+d)d}{2}$ parameters

thus total need $d + \frac{(1+d)d}{2}$ parameters

3. Say we have 3 random variables A, B, and C. Say you are told A is independent of B and that B is independent of C. Does this imply that A is independent of C? Answer yes or no and justify your answer in 1 sentence.



No, just make A equal to C, then A is dependent on C while A is independent of B, B is independent of C

4. Let $f(x, y)$ be a density function for two real-valued random-variables X and Y. Clearly state the conditions that the function f must satisfy to be a density function.

1. $\int_X \int_Y f(x,y)\, dy\, dx = 1$

2. $f(x,y) \geq 0$ for $\forall x \in X$, $\forall y \in Y$

5. Consider a naive Bayes model where we have a variable C taking M possible values $c \in \{1, \ldots, M\}$ and T discrete features $X_1, \ldots, X_T$, each taking K possible values $\{1, \ldots, K\}$. Each feature $X_t, t = 1, \ldots, T$ is conditionally independent of all the other features given C, and C is marginally independent of the X's.

(a) Write down an expression for the joint distribution $P(C, X_1, \ldots, X_T)$ given the information above.

$$P(C, X_1, \cdots, X_t) = P(C \mid Pa(C)) \prod_{i=1}^{t} P(X_i \mid pa(X_i))$$
$$= P(C) \prod_{i=1}^{t} P(X_i \mid C)$$

(b) Specify precisely how many parameters are in this model. Include "-1" terms in your expression.

for $P(C)$ since sum to 1, it has $M-1$ parameters

for each X_i, since $\sum_{d=1}^{K} P(X_i = d \mid C^*) = 1$ and C has M possiblity

then has $M(K-1)$ parameters

in total has $T \cdot M(K-1) + M - 1$ parameters

Problem 2: (20 points)

Let X be a geometric random variable taking values $x \in \{0, 1, 2, \ldots\}$, with probability distribution defined as

$$P(x) = (1-\theta)^x \theta, \quad \text{for } x = 0, 1, 2, \ldots \tag{1}$$

where θ is the parameter of the geometric model and $0 < \theta < 1$. The geometric distribution models a problem where we have a sequence of random binary outcomes with "success" probability θ, where the outcomes are generated independently at each step, and x is the number of failures before success. For example this could be a model of the number of tails we see in tossing a coin before we see heads.

Below we will be using a Beta density as a prior for θ, is defined as

$$P(\theta) = Be(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\alpha, \beta > 0$ and $B(\alpha, \beta) = B(\beta, \alpha)$ is a normalization constant that does not depend on θ. The mean of the prior is $\frac{\alpha}{\alpha+\beta}$ and the mode is at $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha > 1, \beta > 1$.

Let $D = \{x_1, \ldots, x_N\}$ be an observed data set where we assume the samples were generated in an IID manner from $P(x)$.

1. Define the likelihood function for this problem.

$$P(D|\theta) \overset{iid}{=} \prod_{i=1}^{N} p(x_i|\theta)$$

$$= \prod_{i=1}^{N} (1-\theta)^{x_i} \theta$$

$$= \theta^N (1-\theta)^{\sum_{i=1}^{N} x_i}$$

2. Prove that the Beta density $Be(\alpha, \beta)$ is a conjugate prior for the Geometric likelihood and derive an equation for the posterior density for θ.

$$p(\theta|D) \propto p(D|\theta)\, p(\theta)$$

$$= \theta^N (1-\theta)^{\sum_{i=1}^N x_i} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$= \theta^{\alpha+N-1} (1-\theta)^{\sum_{i=1}^N x_i + \beta - 1}$$

WLOG let $\alpha' = \alpha + N$, $\beta' = \sum_{i=1}^N x_i + \beta$

then $p(\theta|D) \propto Be(\alpha', \beta')$

thus $Be(\alpha, \beta)$ is a conjugate prior

3. Define θ^{MAP} for this problem, assuming $\alpha > 1, \beta > 1$, and write a 1 or 2 sentence intuitive interpretation of how this estimate relates to θ^{ML} for the same problem.

$$\theta^{MAP} = \arg\max_{\theta} p(\theta|D)$$

consider $\ell(\theta|D) = (\alpha+N-1)\log\theta + (\sum_{i=1}^N x_i + \beta - 1)\log(1-\theta)$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\alpha+N-1}{\theta} - \frac{\sum_{i=1}^N x_i + \beta - 1}{1-\theta} \overset{want}{=} 0$$

$$\alpha+N-1 - (\alpha+N-1)\theta = (\sum_{i=1}^N x_i + \beta - 1)\theta$$

$$\theta^{MAP} = \frac{\alpha+N-1}{(\alpha+N-1) + (\sum_{i=1}^N x_i + \beta - 1)}$$

$$\ell(\theta) = (\alpha-1)\log\theta - (\beta-1)\log(1-\theta) \Rightarrow \theta^{ML} = \frac{\alpha-1}{(\alpha-1)+(\beta-1)}$$

intutive : θ^{MAP} is the convex linear combination of θ^{ML} and empirical mean $\frac{N}{N+\sum_{i=1}^N x_i -}$

$\alpha-1$

Problem 3: (20 points)

We have two coins. The probability of heads for coin 1 is $0.5 + \phi$ and the probability of heads for the second coin is $0.5 - \phi$, where $0 \leq \phi \leq 0.5$. In words, one coin is biased above 0.5 by some unknown amount and the other coin is biased below 0.5 by the same unknown amount. Say we have data $D_1 = \{x_i\}, i = 1, \ldots, N_1, x_i \in \{0, 1\}$ for N_1 observations from the first coin, and $D_2 = \{x_j\}, j = 1, \ldots, N_2, x_j \in \{0, 1\}$ for the ~~second~~ two coin, where each of the x_i and x_j observations are conditionally independent given the ϕ.

1. Define the likelihood $L(\phi)$ for this problem.

$$
\begin{aligned}
L(\phi) &= P(D_1 \mid \phi) \; P(D_2 \mid \phi) \\
&= \prod_{i=1}^{N_1} p(X_i \mid \phi) \prod_{j=1}^{N_2} P(X_j \mid \phi) \\
&= \prod_{i=1}^{N_1} (0.5+\phi)^{X_i} (0.5-\phi)^{1-X_i} \prod_{j=1}^{N_2} (0.5-\phi)^{X_j} (0.5+\phi)^{1-X_j} \\
&= (0.5+\phi)^{\sum_{i=1}^{N_1} X_i + N_2 - \sum_{j=1}^{N_2} X_j} \; (0.5-\phi)^{N_1 - \sum_{i=1}^{N_1} X_i + \sum_{j=1}^{N_2} X_j}
\end{aligned}
$$

2. Derive the maximum likelihood estimator for ϕ. It may simplify your notation to estimate $\theta = 0.5 + \phi$ and then write $\hat{\phi} = \hat{\theta} - 0.5$.

$$
\begin{aligned}
\ell(\phi) &= \log L(\phi) \\
&= \underbrace{\left(\sum_{i=1}^{N_1} X_i + N_2 - \sum_{j=1}^{N} X_j \right)}_{\text{let = chuck } a} \log (0.5+\phi) + \underbrace{\left(N_1 - \sum_{i=1}^{N} X_i + \sum_{j=1}^{N_2} X_j \right)}_{\text{let = chuck } b} \log (0.5-\phi)
\end{aligned}
$$

$$0 \overset{\text{want}}{=} \frac{\partial \ell}{\partial \phi} = \frac{a}{0.5+\phi} - \frac{b}{0.5-\phi}$$

then $$\frac{a}{0.5+\phi} = \frac{b}{0.5-\phi}$$

$$
\begin{aligned}
0.5a - a\phi &= 0.5b + b\phi \\
0.5(a-b) &= (a+b)\phi \\
\phi^{ML} &= \frac{0.5(a-b)}{a+b} = \frac{2\sum_{i=1}^{N_1} X_i - 2\sum_{j=1}^{N_2} X_j + N_2 - N_1}{2(N_1+N_2)}
\end{aligned}
$$

3. Say you have some prior knowledge that ϕ is roughly 0.1, but your knowledge is relatively weak. Suggest how you might put a prior on ϕ if you wanted to do a Bayesian analysis. No need for detailed equations (unless you wish to): a sentence or two and a simple equation should be fine, as long as it is clear what type of prior you are describing.

this kernel is like the dirchelet distribution when $k=2$

thus we can set a dirchelet prior for $\phi = 0.1$ with ~~same~~ small α_1 and α_2

Problem 4: (20 points)

Let $Z = (X, Y)$ be a two-dimensional (column vector) random variable taking values $\underline{z} = (x, y)$, where x and y are real-valued. Say we have a Gaussian model for $p(\underline{z})$ where

$$p(x, y) = p(\underline{z}) = \frac{1}{C}\exp\Big(-\frac{1}{2}(\underline{z} - \mu_z)^T \Sigma_z^{-1}(\underline{z} - \mu_z)\Big)$$

where C is a normalization constant not involving $\underline{z}$, x, y, and Σ_z is the covariance matrix

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

and where $\mu_z = (\mu_x, \mu_y)$ is the mean (column) vector for Z, T indicates transpose.

* Change notation let $\underline{z} = (X_1, X_2)$ $\mu_z = (\mu_1, \mu_2)$

1. Prove that if Σ_z is diagonal (i.e., $\sigma_{xy} = 0$) then X and Y are independent.

$(\Rightarrow)$ let Σ_z be diagonal $\Sigma_z = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$, $\Sigma_z^{-1} = \begin{pmatrix} \sigma_x^{-2} & \\ & \sigma_y^{-2} \end{pmatrix}$

then $p(x, y) \propto e^{(-\frac{1}{2}(X_1 - \mu_1, X_2 - \mu_2)\begin{pmatrix} \sigma_x^{-2} & \\ & \sigma_y^{-2} \end{pmatrix}\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix})}$

$= e^{-\frac{1}{2}[(X_1 - \mu_1)\sigma_x^{-2}(X_1 - \mu_1) + (X_2 - \mu_2)\sigma_y^{-2}(X_2 - \mu_2)]}$

$= e^{-\frac{1}{2}(X_1 - \mu_1)^2 \sigma_x^{-2}} \quad e^{-\frac{1}{2}(X_2 - \mu_2)^2 \sigma_y^{-2}}$

$\propto p(x)\, p(y)$

thus X and Y are independent

$(\Leftarrow)$ let X, Y independent

$p(x, y) = p(x)\, p(y) \propto e^{-\frac{1}{2}(X_1 - \mu_1)^2 \sigma_x^{-2}} \, e^{-\frac{1}{2}(X_2 - \mu_2)^2 \sigma_y^{-2}}$

$= e^{-\frac{1}{2}((X_1 - \mu_1)\sigma_x^{-2}(X_1 - \mu_1)^t + (X_2 - \mu_2)\sigma_y^{-2}(X_2 - \mu_2)^t)}$

$= e^{-\frac{1}{2}((X_1 - \mu_1, X_2 - \mu_2)\begin{pmatrix} \sigma_x^{-2}(X_1 - \mu_1)^t \\ \sigma_y^{-2}(X_2 - \mu_1)^t \end{pmatrix})}$

$= e^{-\frac{1}{2}(\underline{z} - \mu_z)\begin{pmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{pmatrix}(\underline{z} - \mu_z)^t)}$

Since independent thus $\sigma_{xy} = 0$, $\sigma_{xy} = 0$

thus $\Sigma_z = \begin{pmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_x^2 & \\ & \sigma_y^2 \end{pmatrix}$ is diagonal ∎

2. Say we now have data $D = \{(x_i, y_i)\}$ consisting of ~~conditionally~~ independent samples given the parameters of the Gaussian model for $p(\underline{z})$. The means μ_x and μ_y are known, the variances σ_x^2 and σ_y^2 are unknown, and Σ_z is diagonal. Prove that the log-likelihood $\log L(\sigma_x^2, \sigma_y^2)$ can be written as a sum in the form of $\log L(\sigma_x^2) + \log L(\sigma_y^2)$.

$$L(\sigma_x^2, \sigma_y^2) = P(D \mid \cancel{\mu}, \sigma_x^2, \sigma_y^2) = \prod_{i=1}^{N} (p(x,y) \mid \sigma_x^2, \sigma_y^2)$$

by ①

$$= \prod_{i=1}^{N} p(x \mid \sigma_x^2) \prod_{i=1}^{N} p(y \mid \sigma_y^2)$$

$$\propto \prod_{i=1}^{N} e^{-\frac{1}{2}(x_i-\mu_i)^2/\sigma_x^2} \prod_{i=1}^{N} e^{-\frac{1}{2}(x_i-\mu_i)^2/\sigma_y^2}$$

$$\ell(\sigma_x^2, \sigma_y^2) = \sum_{i=1}^{N} \left(-\frac{1}{2}(x_i-\mu_i)^2/\sigma_x^2\right) + \sum_{i=1}^{N} \left(-\frac{1}{2}(x_i-\mu_i)^2/\sigma_y^2\right)$$

considen $\ell(\sigma_x^2) = \log \prod_{i=1}^{N} p(x \mid \sigma_x^2)\,) = \sum_{i=1}^{N} \left(-\frac{1}{2}(x_i-\mu_i)^2/\sigma_x^2\right)$ same for $\ell(\sigma_y^2)$

thus $\log L(\sigma_x^2) + \log L(\sigma_y^2) = \log L(\sigma_x^2, \sigma_y^2)$

3. Using the information in the first two parts of the problem derive the maximum likelihood estimate for σ_x^2 where the means μ_x and μ_y are known and the variances σ_x^2 and σ_y^2 are unknown.

change $\mu_1 = \mu_x$ $\mu_2 = \mu_y$

$$\ell(\sigma_x^2, \sigma_y^2) = -\frac{1}{2}\sum_{i=1}^{N}\left((x_i-\mu_1)^2/\sigma_x^2 + (x_i-\mu_2)^2/\sigma_y^2\right)$$

by ② drop latter part

$$\frac{\partial \ell(\sigma_x^2, \sigma_y^2)}{\partial \sigma_x^2} \underset{\substack{\text{variable} \\ a=\sigma_x^2}}{\overset{\text{change of}}{=}} \frac{\partial}{\partial a}\log\left(\prod \frac{1}{\sqrt{2\pi a}} e^{-\frac{1}{2}\frac{(x_i-\mu_1)^2}{a}}\right)$$

$$= \frac{\partial}{\partial a}\left[\sum_{i=1}^{N} -\frac{1}{2}\log(2\pi a) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu_1)^2/a\right]$$

$$= \sum_{i=1}^{N} -\frac{1}{2}\frac{2\pi}{2\pi a} + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu_1)^2/a^2$$

$$\overset{\text{want}}{=} 0$$

thus $$\frac{1}{\cancel{2}}\sum_{i=1}^{N}\frac{1}{a} = \frac{1}{\cancel{2}a^2}\sum_{i=1}^{N}(x_i-\mu_1)^2$$

$$a N = \sum_{i=1}^{N}(x_i-\mu_1)^2$$

$$\sigma_x^{2\,ML} = a^{ML} = \frac{\sum_{i=1}^{N}(x_i-\mu_1)^2}{N}$$

Additional Solution Material for Problem `write number here:`

Additional Solution Material for Problem `write number here`**:**