Statistical Machine Learning (W4400)
Spring 2016
https://courseworks.columbia.edu

John P. Cunningham
jpc2181

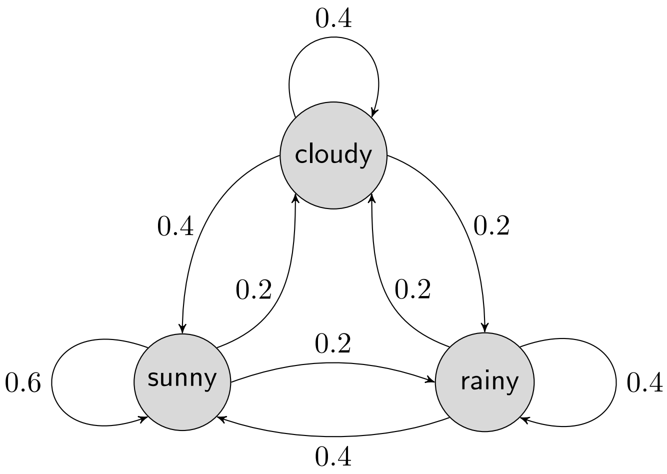
Ben Reddy, Phyllis Wan, Ashutosh Nanda
bmr2136, pw2348, an2655

Practice Final Exam 2

Do not open this exam until instructed. Carefully read the following instructions.

You have 120 minutes to complete the entirety of this exam. Write your name, UNI, and the course title on the cover of the blue book. All solutions should be written in the accompanying blue book. No other paper (including this exam sheet) will be graded. **To receive credit for this exam, you must submit blue book with the exam paper placed inside.** As reference you may use one sheet of 8.5×11in paper, on which any notes can be written (front and back). No other materials are allowed (including calculators, textbooks, computers, and other electronics). To receive full credit on multi-point problems, you must thoroughly explain how you arrived at your solutions. Each problem is divided up into several parts. Many parts can be answered independently, so if you are stuck on a particular part, you may wish to skip that part and return to it later. Good luck.

1. **Weather modeling** (30 points)
A heavily simplified weather model on a sequence of days is captured in the following Markov model (weather is assumed to take on one and only one state in each day):



(a) (5 points) What is the transition matrix of this Markov chain? In all parts of this problem, order the state sequence as $\{ S , R , C \} = \{$ sunny , rainy , cloudy $\}$.

Solution:

$$\begin{bmatrix} 0.6 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.4 \end{bmatrix}$$

(b) (5 points) If today is sunny S, what is the probability of the next three days being rainy R?

Solution:

$$P(R|R)P(R|R)P(R|S) = 0.4 \cdot 0.4 \cdot 0.2.$$

(c) (5 points) If yesterday was sunny S, and tomorrow will be rainy R, what is the most likely weather today?

Solution:
We see X such that $P(R|X)P(X|S)$ is maximized. Consider each possibility individually:

$$\begin{aligned} X = S &\quad \Rightarrow \quad P(R|X)P(X|S) = 0.2 \cdot 0.6 \\ X = R &\quad \Rightarrow \quad P(R|X)P(X|S) = 0.4 \cdot 0.2 \\ X = C &\quad \Rightarrow \quad P(R|X)P(X|S) = 0.2 \cdot 0.2 \end{aligned}$$

and thus the maximum is achieved when today is sunny S.

(d) (2 points) Is this Markov chain irreducible?

Solution:

Yes, fully connected with nonzero probabilities (which is sufficient).

(e) (2 points) Is this Markov chain aperiodic?

Solution:

Yes, each state has a nonzero probability of staying in the same state (which is sufficient).

(f) (5 points) If it exists, what is the equilibrium distribution of this Markov chain?
Hint: the transition matrix p has the form $p = \frac{4}{5}p_1 + \frac{1}{5}I$ for some matrix p_1. Once you find p_1, the answer should be apparent, but you should test your answer to be sure.

Solution:

$$P_{eq} = \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}.$$

(g) (1 point) If it exists, what is the eigenvalue associated with the equilibrium distribution above?

Solution:

$1.$

(h) (2 points) If you have data and wish to fit the transition probabilities with maximum likelihood (rather than using the probabilities given above), how many free parameters will you fit?

Solution:

6; since each column has 2 free parameters (and the other determined by $1 - \sum p_i$).

(i) (2 points) The graphical model above corresponds to a Markov chain of order 1. If you instead wish to fit the transition probabilities of a Markov chain of order 2, how many free parameters must you fit?

Solution:

2 free parameters for each $P(X_i|X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2})$, and there are 9 such distributions (one for each choice of x_{i-1}, x_{i-2}, so 18 total parameters.

(j) (1 point) In maximum likelihood we calculate the likelihood of our data at the best parameter setting. In general, the maximum likelihood for the Markov chain of order 2 will be larger than the maximum likelihood for the Markov chain of order 1, due to the larger number of parameters. What penalization strategies exist to help choose between these two models?

Solution:

AIC and BIC.

2. **Expectation-Maximization** (40 points)

In class we studied Expectation-Maximization in the context of a Gaussian mixture model, and in the homework you programmed a multinomial mixture model. Here you will consider the Poisson mixture model.

Recall the Poisson distribution is an exponential family model, with support over non-negative integers, $X \in \{0, 1, 2, 3, \ldots\}$. The Poisson distribution has a single positive real parameter, $\lambda \in \mathbb{R}_+$. The parameter is both the expected value and the variance of the distribution: $\mathbb{E}(X) = \lambda$ and $Var(X) = \lambda$. The poisson pmf is:

$$p(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Be careful: this problem has part i on the next page. Be sure not to miss it.

(a) (4 points) Write the Poisson distribution in its exponential family form. Specifically, identify the sufficient statistic $S(x)$, the base measure $h(x)$, the natural parameter(s) θ, and the normalizing constant $Z(\theta)$ of the Poisson distribution.

Solution:

$$p(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{\frac{1}{x!}}{e^{\lambda}} \exp\left(x \log \lambda\right) = \frac{h(x)}{Z(\theta)} \exp\left\{S(x)^\top \theta\right\}$$

(b) (2 points) We draw n points iid from a Poisson model with parameter λ. What is the log likelihood, $\mathcal{L}(\lambda)$ of the data?

Solution:

$$\mathcal{L}(\lambda) = \sum_{i=1}^{n} \log p(x_i|\lambda) = \sum_{i=1}^{n} (x_i \log \lambda - \log x_i!) - n\lambda$$

(c) (4 points) Now imagine these n datapoints x_i are drawn from a mixture of K poisson distributions. For each cluster we have a cluster proportion, c_k, where $c_k \geq 0$ and $\sum_{k=1}^{K} c_k = 1$, and a parameter λ_k. Now what is the log likelihood, $\mathcal{L}(\lambda, c)$, of the data?

Solution:

$$\mathcal{L}(\lambda, c) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} c_k p(x_i|\lambda_k) \right)$$

(d) (2 points) Expectation-Maximization can be seen as an optimization problem. What is the function EM is optimizing, is it guaranteed to converge, and if so will it be a local or global optimum?

Solution:

(e) (5 points) What is the E-Step for the Poisson mixture model? Write down the formula to compute the posterior values $P(m_i = k|x_i)$ for step $j + 1$, given the current values at step j. In the slides we called this $a_{ik}^{(j+1)}$.

Solution:

$$a_{ik}^{j+1} := \frac{c_k^{(j)} p(x_i|\lambda_k^{(j)})}{\sum_{l=1}^{K} c_l^{(j)} p(x_i|\lambda_l^{(j)})}$$

(f) (5 points) What is the M-Step for the Poisson mixture model? Write down the formulas to compute $c_k^{(j+1)}$ and $\lambda_k^{(j+1)}$ using the a values at step $(j+1)$ computed in the previous part.

Solution:

$$c_{ik}^{(j+1)} := \frac{\sum_{i=1}^{n} a_{ik}^{(j)}}{n}$$

and

$$\lambda_k^{(j+1)} := \frac{\sum_{i=1}^{n} a_{ik}^{(j)} x_i}{\sum_{i=1}^{n} a_{ik}^{(j)}}$$

(g) (2 points) As with the Gaussian mixture model, this Poisson mixture model can be used as a clustering method. Write down the clustering rule based on the results of running the EM algorithm until it converges.

Solution:

(h) (8 points) A genie gives you 5 data points and says they are from a poisson mixture model with 2 components. Thus we take $K = 2$, and $n = 5$. The genie will reward you if you can correctly identify the cluster each point belongs to. The genie's data is: $x = \{\ 3,\ 8,\ 4,\ 7,\ 5\ \}$. You decide to use EM. First, initialize:

- $c_1^0 = c_1^{(0)} = \frac{1}{2}$
- $\lambda_1^{(0)} = min(x) = 3$
- $\lambda_2^{(0)} = max(x) = 8$

Calculate the E-step for each of the 5 data points. You may want to put your answers in a 5x2 a matrix, as in table **??**.

Table 1: E-Step Answer Template

x_i	a_{i1}	a_{i2}
3		
8		
4		
7		
5		

Note: You may find table **??** helpful.

Note: The genie is very good at arithmetic, so you may leave results as, eg, $\frac{.5\times.37}{(.5\times.37+.5\times.14)}$

Table 2: Poisson Distribution Probability Table

x	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$	$\lambda = 8$	$\lambda = 9$	$\lambda = 10$
0	0.37	0.14	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00
1	0.37	0.27	0.15	0.07	0.03	0.01	0.01	0.00	0.00	0.00
2	0.18	0.27	0.22	0.15	0.08	0.04	0.02	0.01	0.00	0.00
3	0.06	0.18	0.22	0.20	0.14	0.09	0.05	0.03	0.01	0.01
4	0.02	0.09	0.17	0.20	0.18	0.13	0.09	0.06	0.03	0.02
5	0.00	0.04	0.10	0.16	0.18	0.16	0.13	0.09	0.06	0.04
6	0.00	0.01	0.05	0.10	0.15	0.16	0.15	0.12	0.09	0.06
7	0.00	0.00	0.02	0.06	0.10	0.14	0.15	0.14	0.12	0.09
8	0.00	0.00	0.01	0.03	0.07	0.10	0.13	0.14	0.13	0.11
9	0.00	0.00	0.00	0.01	0.04	0.07	0.10	0.12	0.13	0.13
10	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.10	0.12	0.13

eg $p(X = 2; \lambda = 5) = .08$

Solution: We note that since $c_k^{(0)} = .5$ for both clusters it cancels and can be omitted.

x_i	a_{i1}	a_{i2}
3	.22 / (.22 + .03) = .88	.03 / (.22 + .03) = .12
8	.01 / (.01 + .14) = .067	.14 / (.01 + .14) = .933
4	.20 / (.20 + .06) = .77	.06 / (.20 + .14) = .23
7	.02 / (.02 + .14) = .125	.14 / (.02 + .14) = .875
5	.10 / (.10 + .09) = .53	.09 / (.10 + .09) = .47

(i) (8 points) **Calculate the M-step: use the values from your E-Step and the data to update the parameter values: $c_k^{(1)}$, $\lambda_k^{(1)}$.**

Solution:

Let $N_k = \sum_{i=1}^{N} a_{ik}$, which is the effective number of data points assigned to component k (ie the column sums of the a table).

$$N_1 = 2.372, \quad N_2 = 2.628$$

Then, the new mixture weights are

$$c_1^{(1)} = \frac{N_1}{N} = 0.4744, \;\; c_2^{(1)} = \frac{N_2}{N} = 0.5256$$

the updated parameters are

$$\begin{aligned} \mu_1^{new} &= \frac{1}{N_1} \sum_{i=1}^{N} a_{i1} x_i = 0.4123, \\ \mu_2^{new} &= \frac{1}{N_2} \sum_{i=1}^{N} a_{i2} x_i = 6.55 \end{aligned}$$

3. **Categorical Bayesian Models** (30 points)
The categorical distribution is a generalization of the more familiar Bernoulli distribution. Specifically, for K discrete outcomes, and given a K-dimensional parameter vector θ, the distribution of a categorical observation is

$$p(x = k|\theta) = \theta_k = \prod_{j=1}^{K} \theta_j^{\mathbb{1}(x=j)}$$

We will also consider the Dirichlet distribution, with K-dimensional hyperparameter vector α, which we write as:

$$p(\theta) = \frac{1}{D(\alpha_1, ..., \alpha_K)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1},$$

where the normalizer $D(\alpha_1, ..., \alpha_K) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{K} \alpha_j)}$ (though this last detail will not be needed here).

(a) (2 points) Note that the parameter vector θ is a probability mass function on $\{1, ..., K\}$. In class we wrote $\theta \in \Delta_K$, where Δ_K is the K-probability simplex. What specific constraints does the K-probability simplex imply, in terms of θ?

Solution:

$$\theta_j \geq 0, \sum_{j=1}^{K} \theta_j = 1.$$

(b) (2 points) The Dirichlet is a probability density on Δ_K. What then is $\int_{\Delta_K} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} d\theta$?

Solution:

$$D(\alpha_1, ..., \alpha_K).$$

(c) (4 points) You observe n data points drawn conditionally iid, given θ, from the likelihood $p(x|\theta)$ given above. Using the prior $p(\theta)$ given above, write the joint $p(\theta, x_1, ..., x_n)$ and combine terms as much as possible. You may use the notation $z_j \triangleq \sum_{i=1}^{n} \mathbb{1}(x_i = j)$ to denote the count of outcome k.

Solution:

$$\begin{aligned}
p(\theta, x_1, ..., x_n) &= p(\theta) \prod_{i=1}^{n} p(x_i|\theta) \\
&= \frac{1}{D(\alpha_1, ..., \alpha_K)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \prod_{i=1}^{n} \prod_{j=1}^{K} \theta_j^{\mathbb{1}(x_i=j)} \\
&= \frac{1}{D(\alpha_1, ..., \alpha_K)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1 + \sum_{i=1}^{n} \mathbb{1}(x_i=j)} \\
&= \frac{1}{D(\alpha_1, ..., \alpha_K)} \prod_{j=1}^{K} \theta_j^{(\alpha_j + z_j) - 1}.
\end{aligned}$$

(d) (4 points) Is the Dirichlet $p(\theta)$ the conjugate prior to the categorical $p(x|\theta)$? Explain.

Solution: Yes, because the joint distribution has (unnormalized) Dirichlet form.

(e) (5 points) What is the marginal likelihood of those data observations $p(x_1, ..., x_n)$? Hint: be careful here, and be sure to use answers from the previous parts.

Solution:

$$\begin{aligned} p(x_1, ..., x_n) &= \int_{\Delta_K} p(\theta, x_1, ..., x_n) d\theta \\ &= \int_{\Delta_K} \frac{1}{D(\alpha_1, ..., \alpha_K)} \prod_{j=1}^{K} \theta_j^{(\alpha_j + z_j) - 1} d\theta \\ &= \frac{D(\alpha_1 + z_1, ..., \alpha_K + z_K)}{D(\alpha_1, ..., \alpha_K)}. \end{aligned}$$

(f) (4 points) Recall the definition of entropy $H(p(\theta)) = -\int_{\Delta_K} p(\theta) \log p(\theta) d\theta$. This entropy of the prior $p(\theta)$ is simply a function of the hyperparameters $\alpha_1, ..., \alpha_K$, so for simplicity we can write $H(p(\theta) \triangleq h(\alpha_1, ..., \alpha_K)$. For further simplicity assume $n = 1$. Write the conditional entropy of the posterior distribution $p(\theta|X)$ in terms of the function h. Note that this should not simplify much.

Solution:

$$\begin{aligned} H(\theta|X_1) &= \sum_{j=1}^{K} H(q(\theta|X_1 = j)) p(X_1 = j) \\ &= \frac{D(\alpha_1 + 1, \alpha_2 ..., \alpha_K)}{D(\alpha_1, ..., \alpha_K)} h(\alpha_1 + 1, \alpha_2, ..., \alpha_K) + ... + \frac{D(\alpha_1, ..., \alpha_K + 1)}{D(\alpha_1, ..., \alpha_K)} h(\alpha_1, ..., \alpha_K + 1). \end{aligned}$$

(g) (4 points) Is $H(p(\theta|X)) \geq, \leq, =, <,$ or $> H(p(\theta))$? Why?

Solution: The mutual information $I(\theta; X) = H(\theta) - H(\theta|X) \geq 0$, with equality only when X and θ are independent, which in this case they are not. Thus (excluding the degenerate case where some α_i is not finite and thus $H(\theta) = 0$), we know that $H(\theta|X) < H(\theta)$.

(h) (5 points) In class we regularly considered the multinomial likelihood, namely:

$$p(Z|\theta) = \frac{n!}{z_1! ... z_K!} \prod_{j=1}^{K} \theta_j^{z_j}, \quad \text{where } Z = \begin{bmatrix} z_1 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \mathbb{1}(x_i = 1) \\ \vdots \\ \sum_{i=1}^{n} \mathbb{1}(x_i = K) \end{bmatrix}.$$

If we had considered a multinomial likelihood instead of a categorical in all parts above, (how) does this change alter the posterior $p(\theta|Z)$? (How) does this change alter the marginal likelihood $p(Z)$?

Solution: The posterior value $p(\theta|Z)$ is unchanged. The likelihood is scaled by the normalizer of the multinomial. From this we can conclude that the posterior does not consider the order of observations, which makes sense given the conditionally iid generative process.