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 1

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). Also a calculator is allowed for simple calculations. No other materials are allowed (including 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. **Clustering** (25 points)

Suppose you are given the following 1-dimensional data set consisting of four data points: $D = \{x_1, ..., x_4\}$, where

$$\begin{aligned} x_1 &= 0.5, \\ x_2 &= 1, \\ x_3 &= 1.5, \\ x_4 &= 3.5. \end{aligned}$$

Assume that each data point is drawn iid from a Gaussian mixture model with 2 components:

$$p(x|\Theta) = \sum_{k=1}^{2} c_k p(x|M_k, \theta_k = (\mu_k, \sigma_k))$$

where each mixture component is a Gaussian density with mean μ_k and variance σ_k^2, and for each observation x_i, $M_i = \begin{bmatrix} M_{i1} \\ M_{i2} \end{bmatrix} \in \{0,1\}^2$ is the hard assignment vector (a binary vector, just as shown in lecture) indicating the cluster to which data point x_i belongs. Also, c_k are the mixture weights, representing the probability that a randomly selected x_i was generated by component k, $c_k \geq 0, \sum_{k=1}^{2} c_k = 1$. The complete set of parameters for this mixture model with 2 components is thus $\Theta = \{c_1, c_2, \theta_1, \theta_2\}$, where $\theta_k = (\mu_k, \sigma_k)$.

(a) (10 points) E-step: suppose, at iteration 0, the current parameter values are $c_1^{(0)} = c_2^{(0)} = 0.5, \mu_1^{(0)} = 0.5, \mu_2^{(0)} = 2, \sigma_1^{(0)} = \sigma_2^{(0)} = 1$. Write the E-step: calculate all "soft assignment" variables $a_{ik} = p(M_{ik} = 1|x_i, \Theta)$, namely the probability that data point x_i is in cluster k, given data D and current parameters Θ? The following quantities might be useful:

x_i	$p(x_i\|M_1, \theta_1)$	$p(x_i\|M_2, \theta_2)$
0.5	0.4	0.13
1	0.352	0.242
1.5	0.242	0.352
3.5	0.004	0.123

Solution: The soft assignment of data point data point x_i in cluster k is given by

$$a_{ik} = p(M_{ik} = 1|x_i, \Theta) = \frac{c_k p(x_i|M_k, \theta_k)}{\sum_{l=1}^{2} c_l p(x_i|M_l, \theta_l)}$$

which follows directly from Bayes rule. Plug in this equation, we compute each a_{ik} as:

x_i	a_{i1}	a_{i2}
0.5	0.755	0.245
1	0.593	0.407
1.5	0.407	0.593
3.5	0.031	0.969

(b) (10 points) M-step: Now use the membership weights and the data to update parameter values: c_k^{new}, μ_k^{new} and $\sigma_k^{new}, k = 1, 2$.

Solution:
Let $N_k = \sum_{i=1}^{N} a_{ik}$, which is the effective number of data points assigned to component k.,

$$\begin{aligned} N_1 &= 1.786, \\ N_2 &= 2.214 \end{aligned}$$

Then, the new mixture weights are

$$\begin{aligned} c_1^{new} &= \frac{N_1}{N} = 0.447, \\ c_2^{new} &= \frac{N_2}{N} = 0.553 \end{aligned}$$

the updated means are

$$\begin{aligned} \mu_1^{new} &= \frac{1}{N_1}\sum_{i=1}^{N} a_{i1}x_i = 0.946, \\ \mu_2^{new} &= \frac{1}{N_2}\sum_{i=1}^{N} a_{i2}x_i = 2.173, \end{aligned}$$

and the updated sd's are

$$\begin{aligned} \sigma_1^{new} &= \sqrt{\frac{1}{N_1}\sum_{i=1}^{N} a_{i1}(x_i - \mu_1^{new})^2} = 0.518 \\ \sigma_2^{new} &= \sqrt{\frac{1}{N_2}\sum_{i=1}^{N} a_{i2}(x_i - \mu_2^{new})^2} = 1.206 \end{aligned}$$

(c) (5 points) The K-means algorithm can be thought of as a simpler, non-probabilistic alternative to using a Gaussian Mixture Model. What are the two conditions under which Gaussian mixture clustering is reduced to K-means?

Solution: K-means has no explicit notion of cluster covariances. One can reduce Gaussian mixture clustering to K-means if one were to (a) fix a $priori$ all the covariances for the K components to be the identity matrix (and not update them during the M-step), and (b) during the E-step, for each data vector, assign a membership probability of 1 for the component it is most likely to belong to, and 0 for all the other memberships (in effect make a ‘hard decision’ on component membership at each iteration).

2. **Bayesian Models** (30 points)
You have a complicated machine that produces biased coins, where the bias θ of each coin is drawn from a Beta distribution. As a reminder, $\theta \sim$ Beta(α, β) has the following pdf:

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}\{0 \leq \theta \leq 1\},$$

where the normalizing constant $B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$ is called the *beta function* (but note that in none of the following questions will you have to perform any operations on the beta function). To help, we note that the Beta distribution has mean and variance:

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \qquad\qquad Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The machine generates a single coin, which you then flip repeatedly. Call the results of these flips x_i, which are conditionally iid Bernoulli, namely $x_i|\theta \sim$ Bern(θ). You observe that $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$. The following problems investigate what you can learn about θ from these observations.

(a) (4 points) Write the Beta distribution in its exponential family form $p(\theta) = \frac{1}{Z(\eta)} \exp \eta^\top S(\theta)$. Note that the natural parameters η are a function only of α and β. You may assume the support is given (that is, you may ignore the term $\mathbb{1}\{0 \leq \theta \leq 1\}$).

Solution:

$$\begin{aligned} p(\theta) &= \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)} \exp\left\{(\alpha-1)\log\theta + (\beta-1)\log(1-\theta)\right\} \\ &= \frac{1}{B(\alpha, \beta)} \exp\left\{\begin{bmatrix}\alpha-1 \\ \beta-1\end{bmatrix}^\top \begin{bmatrix}\log\theta \\ \log(1-\theta)\end{bmatrix}\right\}, \end{aligned}$$

which is the exponential family form. $Z(\eta)$, η, and $S(\theta)$ can be read off directly.

(b) (4 points) What is the likelihood $p(x_1 = 1, x_2 = 1, x_3 = 0|\theta)$?

Solution: This is the joint distribution over three conditionally independent Bernoulli r.v.'s. Thus $p(x_1 = 1, x_2 = 1, x_3 = 0|\theta) = p(x_1 = 1|\theta)p(x_2 = 1|\theta)p(x_3 = 0|\theta) = \theta^2(1-\theta)$.

(c) (6 points) What is the posterior $p(\theta|x_1 = 1, x_2 = 1, x_3 = 0)$?

Solution: To solve this problem we can either exploit the conjugacy of Beta and Bernoulli, or equivalently we can use Bayes Rule to calculate the posterior

$p(\theta|x_1 = 1, x_2 = 1, x_3 = 0) =$

$$\begin{aligned}
&= \frac{p(\theta|x_1 = 1, x_2 = 1, x_3 = 0)p(\theta)}{\int_0^1 p(\theta'|x_1 = 1, x_2 = 1, x_3 = 0)p(\theta')d\theta'} \\
&= \frac{\theta^2(1-\theta)\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}(0 \le \theta \le 1)}{\int_0^1 \theta'^2(1-\theta')\frac{1}{B(\alpha,\beta)}\theta'^{\alpha-1}(1-\theta')^{\beta-1}d\theta'} \\
&= \frac{\theta^2(1-\theta)\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}(0 \le \theta \le 1)}{\int_0^1 \theta'^2(1-\theta')\theta'^{\alpha-1}(1-\theta')^{\beta-1}d\theta'} \\
&= \frac{\theta^{(\alpha+2)-1}(1-\theta)^{(\beta+1)-1}\mathbb{1}(0 \le \theta \le 1)}{\int_0^1 \theta'^{(\alpha+2)-1}(1-\theta')^{(\beta+1)-1}d\theta'} \\
&= \frac{1}{B(\alpha+2,\beta+1)}\theta^{(\alpha+2)-1}(1-\theta)^{(\beta+1)-1}\mathbb{1}(0 \le \theta \le 1)
\end{aligned}$$

where the last line comes from the denominator being another Beta integral (see previous parts). From this we see $\theta|x_1 = 1, x_2 = 1, x_3 = 0 \sim Beta(\alpha + 2, \beta + 1)$, which is indeed the conjugacy relationship between the Beta and Bernoulli distributions.

(d) (4 points) Compare the posterior mean $E(\theta|x_1 = 1, x_2 = 1, x_3 = 0)$ to the prior mean $E(\theta)$. Specifically, for what values of α and β will $E(\theta|x_1 = 1, x_2 = 1, x_3 = 0) \ge E(\theta)$?

Solution: Exploiting the results of the previous parts, $E(\theta|x_1 = 1, x_2 = 1, x_3 = 0) = \frac{\alpha+2}{\alpha+\beta+3}$. Then:

$$\begin{aligned}
\frac{\alpha+2}{\alpha+\beta+3} &\ge \frac{\alpha}{\alpha+\beta} \\
&\Rightarrow \\
2\beta &\ge \alpha
\end{aligned}$$

(e) (4 points) Interpret (in words) values of α and β such that the posterior and prior means are equal: why does this result make sense, given the observed data?

Solution: The posterior and prior means are the same when $2\beta = \alpha$, or rather $E(\theta) = \frac{2}{3}$. This results makes perfect sense with the data, as we have observed data in precisely that ratio: $\frac{2}{3}$ heads, $\frac{1}{3}$ tails, and thus it is reasonable that our posterior mean belief in the bias has not changed from our prior belief. Our confidence has grown (the variance has reduced), but the mean is unchanged.

(f) (4 points) What influence has the data had on your belief of the bias of the coin that you are flipping, for the following cases? Calculate and compare the prior and posterior means.

1. $\alpha = \beta = 1$.
2. $\alpha = \beta = 100$.

Solution:

1. The prior mean $\frac{1}{2}$ has increased to a posterior mean of $\frac{3}{5}$. Indeed $2\beta > \alpha$ in this case, so we expect the posterior mean to be larger.

2. The prior mean $\frac{1}{2}$ has increased to a posterior mean of $\frac{102}{203}$. Indeed $2\beta > \alpha$ in this case, so we expect the posterior mean to be larger, but notice that the change was significantly less than in the first case.

(g) (4 points) Interpret (in words) the results of the previous part (part f) in words. Why do these make sense? Hint: consider the prior variance of the Beta distribution.

Solution: Note that the prior variance of the first case is $\frac{1}{12} \approx 0.083$ and of the second case is much smaller, ≈ 0.0012. The prior uncertainty is thus larger in the first case than in the second, so it makes sense that the data should more significantly influence our posterior belief when our prior is more uncertain (the first case) than when our prior belief is more certain (the second case).

3. **Loss Functions** (20 points)

Throughout this question consider y as a binary variable, $y \in \{-1, 1\}$. The following are loss functions that are frequently encountered in machine learning:

- 0-1 Loss: $L_{01}(y, f(x)) = \begin{cases} 0 & yf(x) > 0 \\ 1 & yf(x) \leq 0. \end{cases}$
- Hinge Loss (used in SVM): $L_h(y, f(x)) = \max\{0, 1 - yf(x)\}$.
- Square Loss: $L_{sq}(y, f(x)) = (1 - yf(x))^2$.
- Exponential Loss (used in boosting): $L_{exp}(y, f(x)) = \exp\{-yf(x)\}$.
- Log Loss (used in logistic regression): $L_{log}(y, f(x)) = \log(1 + \exp\{-yf(x)\})$.

Importantly, notice that all of these loss functions can be written in terms of the margin $z = yf(x)$, such that $L_{sq}(y, f(x)) = L_{sq}(z) = (1 - z)^2$ (and similar for every other loss listed above).

(a) (12 points) Draw (roughly) the shape of all five of these loss functions, with the margin $z = yf(x)$ on the horizontal axis, and the loss $L(z)$ on the vertical axis. Clearly label each line on your graph.

Solution: This is more of less straightforward, only thing to note is that the squared loss is not monotonically decreasing in terms of the margin

(b) (4 points) The previous loss functions can all be thought of as convex approximations to the 0-1 loss function. Looking at a plot of the previous loss functions which one appears intuitively to be the worst approximation to the 0-1 loss function? Which one appears to be the best?

Solution: The Hinge Loss is the best approximation, whereas the squared loss is the worst

(c) (4 points) Consider just the 0-1 loss, the hinge loss, and the exponential loss. Rank the loss functions from highest to lowest in terms of robustness to misspecification of class labels in the data (for example, if the labels on y were randomly switched). List the most robust first and the least robust third. Hint: consider the amount of loss assigned to a point z that is large and negative; this point corresponds to a large margin error (or misspecification).

Solution: The 0-1 loss is most robust, followed by the hinge loss and then the exponential loss

4. **Principal Component Analysis** (25 points)

Let $x \in \mathbb{R}^3$ be a multivariate Gaussian random vector, with distribution $x \sim \mathcal{N}(x; \mu, \Sigma)$, where the mean vector μ and covariance matrix Σ are:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \Sigma_0 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In the following you can assume you have an infinite number of data points, such that the empirical mean and covariance of your data is exactly that of the distribution as given above.

(a) (4 points) What is $v_1 \in \mathbb{R}^3$, the first principal component of this data?

Solution: $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, which can be read off from the covariance matrix or calculated.

(b) (4 points) What percentage of total data variance is captured by projecting the data onto the first principal component?

Solution: The answer is 50%. You can do this two ways, the first is to solve for the eigenvectors of the covariance matrix , and see the variance of the first principal component is 3, for the second it is 2, for the third it is 1. The second way uses the eigendecomposition of the covariance matrix, from which we see the variances of the principal components are the eigenvalues of the covariance matrix

(c) (3 points) We now change the data such that it has covariance matrix:

$$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Does it make sense to run PCA in this situation if our goal is dimensionality reduction?

Solution: We have that the variances of every dimension is 1, as these are the eigenvalues of the covariance matrix. It does not make sense to run principal components in this case.

(d) (2 points) We now change the data such that it has covariance matrix:

$$\Sigma_2 = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}.$$

ρ_{ij} can not be arbitrary values. What constraints exist?

Solution: $\Sigma_2 > 0$, to be a valid covariance (in this case correlation) matrix. In other words, the ρ values must be such that the matrix is symmetric and positive definite.

(e) (4 points) In this new data (comparing Σ_2 to Σ_1), the direction of the principal components *or* the total variance of the data has changed. Which? Argue in words.

Solution: The directions of the principal components change, but the total variance does not. This can be verified by completing principal component analysis, or noted by observing that the covariance ellipsoid has simply rotated in $\mathbb{R}^3$.

(f) (8 points) We change the data such that it has covariance matrix:

$$\Sigma_3 = \begin{bmatrix} 1 & \frac{\rho_{12}}{w} & \frac{\rho_{13}}{w} \\ \frac{\rho_{21}}{w} & 1 & \frac{\rho_{23}}{w} \\ \frac{\rho_{31}}{w} & \frac{\rho_{32}}{w} & 1 \end{bmatrix},$$

for some value $w > 1$. Prove that the principal components of this data and the previous data are the same; compare the eigenvectors of Σ_3 and Σ_2. Note: this is a difficult question with a clean analytical answer. We recommend you leave this part until the end of the exam.

Solution: The quickest way to solve this problem is to write $\Sigma_3 = (1 - w^{-1})I + w^{-1}\Sigma_2$. Then we have: $[(1-w^{-1})I + w^{-1}\Sigma_2]v = (1-w^{-1})v + \frac{\lambda v}{w} = (\lambda w^{-1} + (1-w^{-1}))v$, where v is the eigenvector of Σ_2, and λ is the eigenvalue of Σ_2. We see from the above that Σ_3 has the same eigenvector v, and eigenvalue: $\lambda' = \left(1 + \frac{\lambda - 1}{w}\right)$.

Note: one can also write the characteristic polynomial and try to solve for $|\Sigma_3 - \lambda' I|$, which in fact will result in the same solution as above.