To find the posterior density $p(\lambda|\mathcal{X})$ we need a prior on λ. We claim that a conjugate prior for the exponential distribution is the gamma distribution

$$Gamma\,(\lambda|\alpha,\beta) = \frac{\Gamma(\alpha)}{\beta^{\alpha}}\lambda^{\alpha-1}\exp(-\lambda\beta),$$

where $\Gamma(\alpha) = \int_0^{\infty} \exp^{-t}\, t^{\alpha-1}dt$ is the gamma function.

b) 1. What does *conjugate prior* mean? **1 pt.**

it means the posterior would have the same form as the prior

2. Show that the gamma distribution is the conjugate prior of the exponential distribution. **2 pts.**

$p(\lambda|X) \propto p(X|\lambda)\,p(\lambda|\alpha,\beta) \propto \lambda^{n} \exp^{-\lambda(x_1+\cdots x_n)} \cdot \lambda^{\alpha-1} \exp^{-\lambda\beta}$

$\propto \lambda^{\alpha+n-1} \exp^{-\lambda(\beta+x_1+\cdots x_n)}$

which is also a gamma function with $\alpha' = \alpha+n$

$\beta' = \beta + x_1+\cdots x_n$

c) Given a Gamma prior over the rate λ (prior with parametes α and β), write the maximum a posteriori estimator $\widehat{\lambda}_{\text{MAP}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^{n}$: (please write the direct closed-form solution)

$\widehat{\lambda}_{\text{MAP}}(\mathcal{X}) = \arg\max p(\lambda|\mathcal{X}) =$ **2 pts.**

$\log p(\lambda|X) \propto (\alpha+n-1)\log\lambda - \lambda(\beta+x_1+\cdots x_n)$

$\frac{\partial \log p(\lambda|X)}{\partial \lambda} \propto \frac{\alpha+n-1}{\lambda} - (\beta+x_1+\cdots x_n) = 0$

$\Rightarrow \lambda = \frac{\alpha+n-1}{\beta+x_1+\cdots x_n}$

d) When is the maximum likelihood estimatior (MLE) equal to the maximum a posteriori (MAP) estimator given a set of i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^n$?

1. If the number of observations is finite. **2 pts.**

$\frac{n}{x_1+\cdots x_n} = \frac{\alpha+n-1}{\beta+\cdots x_1+\cdots x_n}$ when this equation is is right,

2. If the number of observations is infinite ($n \to \infty$). **2 pts.**

according to the convergence of MLE, in this case it will automatically

$\lim_{n\to\infty} \frac{n}{x_1+\cdots x_n} = \frac{\alpha+n-1}{\beta+\cdots x_1+\cdots x_n}$ in this case $\lambda = \frac{n}{x_1+\cdots x_n}$ $\beta = \alpha\lambda - \lambda$

e) Assume that you have a set of the i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^n$ and that you can not decide which distribution to use for data description: the Gaussian distribution or the Beta distribution. If you use the Bayesian framework what you can look at? **2 pts.**

$p(\lambda|x) = \frac{p(x|\lambda)(p(\lambda))}{\cdots\cdots}$ (prior) compute the likelihood function and the value for the MLE.

Now consider a binary classification task from a set of the i.i.d. observations $\mathcal{X} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, with $\mathbf{x} \in \mathbb{R}^D$. Assume that the likelihood of both classes is Gaussian (assume class prior π_i, mean μ_i, and covariance matrix Σ_i for class y_i, with $i = 1, 2$).

f) Recall that a discriminant function for class y_i is defined as:

$$g_{y_i}(\mathbf{x}) = p(y_i|\mathbf{x}).$$

How can you find a decision surface in terms of likelihood, prior and evidence? **1 pt.**

$p(y_1|x) = p(y_2|x)$

$p(x|y_1)\,p(y_1) = p(x|y_2)\,p(y_2)$

$p(x|y_1)\,p\,\pi_1 = p(x|y_2)\,\pi_2$

4

取 $\ln$: $\frac{1}{2\sigma^2}\|x-\mu_1\|^2 +$

g) Assume that

$$\mu_1 = \mu_2 = \mu$$
$$\Sigma_1 = \frac{1}{2\lambda_1}\mathbb{I}, \quad \Sigma_2 = \frac{1}{2\lambda_2}\mathbb{I}$$
$$\lambda_1 > 0, \quad \lambda_2 > 0, \quad \lambda_1 \neq \lambda_2$$

where $\mathbb{I}$ denotes the identity matrix. Write the equation satisfied by the separating decision surface. The equation must be an explicit function of $\mathbf{x}_1$ (the single observation), of the class prior, means and covariance:
(please write the solution in the polinomial form)

1. Decision surface: **3 pts.**

$g_y(x) = \log P_y(x) + \log \pi_y = -\lambda_y \|x-\mu_y\|^2 + \log \pi_y + \frac{d}{2}\log 2\lambda_y$

$g_{y_1}(x) = g_{y_2}(x)$

$\lambda_1 \|x\|^2 + 2\mu_1 \cdot x - \|\mu_1\|^2 + \log \pi_1 + \frac{d}{2}\log \lambda_1 = \lambda_2 (-\|x\|^2 + 2\mu_2 x - \|\mu_2\|^2) + \log \pi_2 + \frac{d}{2}\log \lambda_2$

$x_1 = \pm \sqrt{\frac{\log \frac{\pi_1 \lambda_1^{d/2}}{\pi_2 \lambda_2^{d/2}}}{\lambda_2 - \lambda_1}}$ $\quad (x-\mu)^2 = \frac{-\log \frac{\pi_1 \lambda_1^{d/2}}{\pi_2 \lambda_2^{d/2}}}{\lambda_2 - \lambda_1}$

2. In the case described above, is the decision surface linear, parabolic, spherical, cylindrical, or something else?

linear|parabolic|spherical|cylindrical|other **2 pts.**

圆柱

depend on the covariance matrix and prior

Question 2: Linear Classifiers and Kernels (20 pts.)

a) Below is a list of algorithms which given a training set output a prediction function. Cross **all** of the algorithms that necessarily output a linear (in the original space) prediction function.

- ☐ Neural network
- ☑ Perceptron with learning rate $\eta = 1$
- ☐ SVM with radial basis kernel
- ☐ K-nearest neighbor classifier
- ☑ SVM with polynomial kernel with degree 1
- ☑ Ridge regression

3 pts.

b) Recall the SVM problem. As a constrained optimization problem a solution can be obtained through both the primal and the dual form.

1. Given a primal solution for the SVM, write down the resulting classifier. **1 pt.**

$g(x) = \sum_{i=1}^{n} \alpha_i z_i \vec{x_i}^T x + w_0$

$w_0 = -\frac{1}{2}\left[\min_{z_i=1} w^{*T} x_i + \max_{z_i=-1} w^{*T} x_i\right]$

$\sum \alpha_i z_i = 0$

$y = \text{sgn}(w^T x + w_0)$

2. Given a dual solution for the SVM, write down the resulting classifier. **1 pt.**

$g(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i^* z_i x_i^T x + w_0\right)$

$\sum \alpha_i^* z_i = 0$

α_i^* is the optimal solution

to $W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \alpha_i \alpha_j y_i^T y_j$

$\forall \alpha_i \geq 0$

3. In practice, often the dual form of the SVM is solved to obtain a classifier. Provide one advantage of solving the dual SVM instead of the primal. **3 pts.**

since the solution to dual SVM induce sparsity to the decision function, where only few support vector enter the sum to evaluate the function

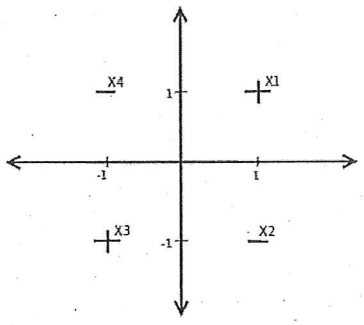
c) Let $S = \{\mathbf{x}_i, y_i\}_{i=1}^4$ be the following training set — efficiency and interpretability apply kernel tricks

$\mathbf{x}_1 = (1, 1)$ $y_1 = 1$,

$\mathbf{x}_2 = (1, -1)$ $y_2 = -1$,

$\mathbf{x}_3 = (-1, -1)$ $y_3 = 1$,

$\mathbf{x}_4 = (-1, 1)$ $y_4 = -1$



Suppose that we trained an SVM on S and the resulting classifier $f(x)$ achieved zero training error. We ask you to provide an explicit description of $f(x)$ (a formula with numeric values).
Hint: Think of a suitable kernel function or alternatively a feature map. **5 pts.**

$(x_1, x_2) \Rightarrow x_1 \cdot x_2$

$x_1 x_1' x_2 x_2'$

$\left(\sum \alpha_i z_i \phi(x_i)^T \cdot \phi(x) + w_0 \right)$

$k = x x' + 9 x_1^2 x_1' x_2 x_2' + x_2 x_2'$

$\phi(x) = [x_1, x_1 x_2, x_2]^T$

$\phi(x) = (1, 1, 1)^T$

$\phi(x_2) = (1, -1, -1)^T$

$\phi(x_3) = (-1, 1, -1)^T$

$\phi(x_4) = (-1, -1, 1)^T$

$\alpha_1 + \alpha_3 = \alpha_2 + \alpha_4$

let $\alpha_1 = \alpha_2 = 1$

$\alpha_3 = \alpha_4 = 2$

$w^* = \sum_{i=1}^{n=4} \alpha_i z_i \phi(x_i)$

$= (1,1,1)^T + (-1)(1,-1,-1)^T + 2(-1,1,-1)^T + 2(-1,-1,1)^T$

$= (-1, 3, 3) + 0(-1, 3, -3)$

$\sum \alpha_i^* z_i = 0$

$= \frac{1}{2}\left((1,1,1)(1,1,1)^T + (1)(1) \right)$

$= \frac{1}{2}(3 + 2)$

$= 0$

$= (0, -2, 2)^T$

$+ 2 + 2 = 6$

or $(x_1, 3 x_1 x_2)$

$k = x_1^2 + 9 x_1 x_1' x_2 x_2'$

d) Consider a training set $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, 1\}$

1. Briefly describe a leave one out (LOO) procedure for estimating the error of an SVM classifier on S. **2 pts**

① make n subset of S, each S_i exclude $[x_i, y_i]$ in training set and make it the validation set C_i,

② train SVM, for each S_i, aim to minimize the training loss (eg, misclassification error)

③ compute the error for, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[C_i(x_i) \neq y_i]$

2. What is the LOO error? **1 pts**

$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[C_i(x_i) \neq y_i]$ or $\frac{1}{n}\sum_{i=1}^{n} (y_i - C_i(x_i))^2$

3. Suppose that we trained an SVM classifier on the **entire** dataset S, denote by sv the set of support vectors, $sv = \{\mathbf{x}_j | \alpha_j > 0\}$.
For the same value of C, prove that the LOO error is bounded by $\frac{|sv|}{n}$ i.e.

$$\text{LOO error} \leq \frac{|sv|}{n}$$

4 pts.

$\text{Sign}\left(\sum_i \alpha_i z_i x_i \times + b\right)$

$R^{CV} = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{I}(C(x_i) \neq y_i)]$ $\frac{1}{n}\#[SV]$ 个.

when $x_i \notin SV$

we have

$\mathbb{I}[C_i(x_i) \neq y_i] = \mathbb{I}(C(x_i \neq y_i)] = 0$

when $x_i \in SV$

$\mathbb{I}[C(x_i) \neq y_i] \leq 1$

Question 3: Bagging and Boosting (20 pts.)

a) Answer precisely the following questions.

1. Are bagging and Boosting Bayesian approaches? Why? **1 pt.**

no, they are heuristic approaches,

2. How is it possible to detect outliers with AdaBoost? **1 pt.**

those who possess large weight ~~are~~ with the interation of the computing progress

3. From the frequentist perspective, bagging is motivated by the tradeoff between two terms. Which? **1 pt.**

Variance and bias

$$\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) - \frac{\sum_{b=1}^{B}\hat{f}_b(x)}{B}\right)^2 = \frac{\sum_{b=1}^{B}(\hat{f}_b(x) - E\hat{f}_b(x))^2}{B^2} = \frac{\sigma^2}{B}$$

4. AdaBoost has an alternative interpretation which is based on the minimization of a certain cost function. Which function? **1 pt.**

$E(e^{-yf(x)})$

5. AdaBoost aims at selecting the best approximation to which ratio? **2 pt.**

error rate $e_b = \frac{\sum_{i=1}^{n} w_i^{(b)} \mathbb{I}(a(x_i) \neq y_i)}{\sum_{i=1}^{n} w_i^{(b)}}$

$\log \frac{1-e_b}{e_b}$

$f(x) = \frac{1}{2} \log \frac{P(y=1|x)}{P(y=-1|x)}$

$\sum_w (\alpha y(x)|x)$

$\sum_w -\alpha y(x))$ $\quad e^{-\alpha C(x)}$

6. Why is the standard form of AdaBoost limited to binary classification?

$G_b = \frac{\sum_{i=1}^{n} w_i^b I(y_i \neq C_b(x_i))}{\sum_{i=1}^{n} w_i^b}$

Since the form of $\sum_w (\alpha y(x)|x)$ correspond to the binary classification. **1 pt.**

7. How could one parallelize bagging?

Since bagging process different Bootstrap dataset independently, so all single classifier could be trained at the same time parallelizely, and the ensemble classifier just run the average. **1 pt.**

8. Name a design property of the base classifiers of AdaBoost which impacts the overall predictive power.

weak classifier

the prediction accuracy should not be too high, and should be higher than 0.5. **1 pt.**

B.

9. Under which conditions does AdaBoost yield good results even when the base classifiers exhibit an individual performance that is only slightly better than that purely due to chance? **2 pt.**

no too much outlier

uncorrected different classifier

B much larger than 1

and large dataset

b) Consider *bagging* in the context of binary classification. Let the *target function* be $h(x)$, where $h : \mathbb{R}^d \to \{\pm 1\}$. Let us combine B *individual classifiers* $y_b(x), b = 1 \ldots B$ to obtain a *committee model*

$$y_{\mathsf{COM}}(x) = \text{sign}\left[\frac{1}{B}\sum_{b=1}^{B} y_b(x)\right]. \qquad (1)$$

1. Write down the pseudocode of bagging for binary classification, from the input (data and model) to the prediction output. **3 pts.**

① bootstrap original dataset to produce B Bootstrap dataset for training
$Z_{b1}, Z_{b2} \cdots Z_{bB}$.
for $i = 1$ to B
② train individual classifiers $y_{bi}(x)$ on Z_{bb} seperately
③ combine the classifier and add equal weight to them, to form the committee model
$y_{com}(x) = \text{sign}\left[\frac{1}{B}\sum_{b=1}^{B} y_b(x)\right]$

2. The error $\epsilon_b(x) = \exp\{-h(x)y_b(x)\}$ indicates the error of an individual model $y_b(x)$ for a single sample x in terms of the target function $h(x)$ and the output of the individual model $y_b(x)$. Write down E_{AV}, that is the average of the expected errors over the individual classifiers $y_b(x)$, and the expected error E_{COM} made by combined model $y_{\mathsf{COM}}(x)$ as a function of the output of the committee model and of the target function.

$E_{\mathsf{AV}} =$ $\frac{\sum_{b=1}^{B} E\{\exp[(-h(x)y_b(x)]\}}{B}$

$E_{\mathsf{COM}} =$ $E\{\exp[-h(x)\,\text{sign}[\frac{1}{B}\sum_{b=1}^{B} y_b(x)]]\}$ **1 pts.**

3. Under which conditions is $E_{\text{AV}} < E_{\text{COM}}$?

$$\frac{\frac{1}{B}\sum_{b=1}^{B} e^{y_b(x)}}{B} < e^{\text{sign}\left[\frac{1}{B}\sum_{b=1}^{B} y_b(x)\right]}$$

3 pts.

4. With the same exponential error, write down the error function for each iteration of AdaBoost with weighting coefficients α_b for the B base classifers $y_b(x)$.

$E_{\text{AdaBoost}} =$ $e^{-y(F^{(b)}(x) + \alpha_b y_b(x))}$

1 pt.

5. In this scenario, within AdaBoost the minimization of this error function is performed with respect to two terms, which?

1) $y_b(x) = \begin{cases} 1 & p_w(y=1|x) > p_w(y=-1|x) \\ -1 & \text{otherwise} \end{cases}$

2) $\alpha_b = \frac{1}{2}\ln\frac{1-\epsilon_b}{\epsilon_b}$

1 pt.

Question 4: Regression, Bias and Variance (20 pts.)

a) Write down the linear regression model (component-wise and in vector notation) for input variable $x = (1, x_1, \ldots, x_D)^\mathsf{T} \in \mathbb{R}^{D+1}$ and output variable y. Formally introduce the model parameter(s).

$y = x^T\beta + \epsilon$

$\epsilon \in \mathbb{R}^{D+1}$ $\quad \beta$ $\quad \epsilon$ mit

From now on, assume that the input dataset consists of N samples given by the matrix $\mathbf{X} \in \mathbb{R}^{N\times(D+1)}$ (where the first column is $\mathbf{1}$), and the output, $\mathbf{Y} \in \mathbb{R}^N$. Write down the linear regression model for all observations $\mathbf{Y}$ (in matrix notation).

2 pts.

$Y = X\beta + \epsilon$

b) Write down the following cost functions (in a notation of your choice): **2 pts.**

1. Ridge Regression (*RR*):

$\sum_{i=1}^{D} (y_i - \sum_{j=1}^{D+1} x_{ij}^T \beta_j)^2 + \lambda \sum_{j=2}^{D+1} \beta_j^2$

$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$

2. Least Absolute Shrinkage and Selection Operator (*LASSO*). Formulate as a constrained optimization problem:

$\min \quad (Y - X\beta)^T(Y - X\beta) \qquad \min_b \sum_{i=1}^{D} (y_i - \sum_{j=1}^{D+1} x_{ij}^T \beta_j)^2$

with $\quad \sum_{j=2}^{D+1} |\beta_j| \leq s$

subject to

c) Formulate the objective of **learning** (formula) in the regression setting introduced above with data $(\mathbf{X}, \mathbf{Y})$ i.i.d. from $P(\mathbf{X}, \mathbf{Y})$.

2 pts.

$\beta = (X^T X)^{-1} X^T Y$

d) Assuming that the observations in $\mathbf{Y}$ are affected by additive Gaussian noise ϵ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. What do we know about the distribution of the *RSS*-estimator (i.e. $\hat{\beta}^{RSS}$)?

3 pts.

$E(a^T\hat{\beta}) = E[a^T (X^T X)^{-1} X^T Y] = a^T [X^T X)^{-1} X^T E(X\beta + \epsilon)$

$= a^T \beta$

$V(a^T\hat{\beta}) = V[a^T (X^T X)^{-1} X^T (X\beta + \epsilon)]$

$= V(a^T\beta) + V(a^T (X^T X)^{-1} X^T \epsilon)$

$= 0 + E(a^T (X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1} a)$

$= \sigma^2 a^T (X^T X)^{-1} a$

$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$

e) Please briefly give another motivation for the *RSS* cost function.

1 pt.

according to Bayesian theory

assume the residual $\epsilon \sim N(0, \sigma^2)$

in order to maximize the like hood

$\max \; P(Y|X,\beta) \propto e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta)}$

$\min (Y - X\beta)^T (Y - X\beta)$

if give $P(\beta|\Lambda) \sim N(0, \Lambda^{-1})$

to maximize the posterior of β

$P(\beta|\Lambda, X, Y) \sim e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta) - \frac{1}{2}\beta^T \Lambda \beta}$

$\min \; (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$

Question 5: Unsupervised Learning (20 pts.)

1. In this section we study non-parametric density estimation of an arbitrary point x. We consider some small region $\mathcal{R}$ containing x. In the class we have seen the following generic formula for density estimation:

$$p(x) = \frac{K}{nV},$$

where K denotes the number of data points falling inside the region $\mathcal{R}$ and V shows the volume of the region. n is the number of data points in the sample set $\mathcal{S} = \{x_1, ..., x_n\}$.

(a) Consider the following Gaussian distribution to be used as a Parzen window function:

$$\phi(x - x_j) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{(x - x_j)^2}{2}\right). \quad (7)$$

What are K and V for this window function?

$K = \sum_i \phi(x - x_i)$

$V = 1$

4 pts.

(b) This particular choice of a window function leads to under-fitting. Add a parameter to increase the model complexity.

too smooth, add h_n

such $p(x) = \frac{1}{n} \sum_i \frac{1}{v_n} \phi\left(\frac{x - x_i}{h_n}\right)$

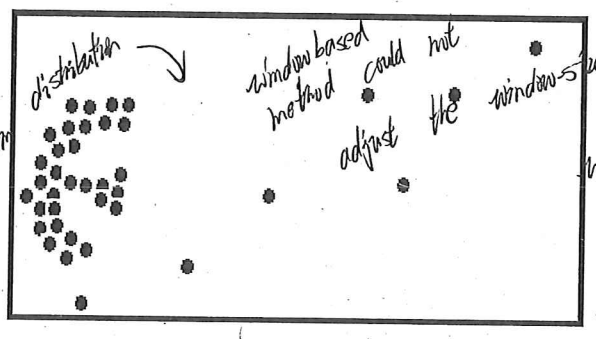
$v_n = (h_n)^d$

2 pts.

let h_n decrease, in this case.

the model will become more complex.

(c) Consider the following sample set. Which of the density estimation methods would you choose? Window-based (Eq. (7)) or K-nearest neighbor? Explain your answer.



Window-based

since it is obvious the data on the right is more sparse than the left one, ~~which~~ thus possessing somehow different densities, distribution,

K-NN ~~was~~ could not take the density difference into account ~~well~~, as good as Window-based is a way of hard assign so we choose the Window-based, though it is still a non-parametric way, each window adds up to better to distinguish the boundry

3 pts.

(d) For a general Parzen window function prove that it provides a probability distribution.

for a Parzen window

it meet the requirement

that $\phi(x) > 0$

$$\int \phi(x)\,dx = 1$$

4 pts.

2. We consider a mixture of K poisson distributions and perform the Expectation-Maximization (EM) algorithm to compute the unknown parameters. The log-likelihood function of n independent objects for mixture of K Poisson distribution is defined as:

$$P(x;\lambda) = \sum_{i=1}^{n} \log \sum_{c=1}^{K} \pi_c f(x_i;\lambda_c)$$

where π_c's are the mixture weights and λ_c's are the parameters of K Poisson distributions. $f(x_i;\lambda_c)$ is defined as:

$$f(x;\lambda_c) = \frac{\lambda^x e^{-\lambda_c}}{x!}.$$

(a) Introduce the latent indicator variables necessary for maximizing the log-likelihood function.

m_{ic} which equals 1 when x is assign to class c

otherwise 0

$P(m_{ic}=1) = \pi_c$

$P(x_n;\pi) = \prod_{k=1}^{K} \pi_k^{m_{ic}}$ **2 pts.**

(b) Calculate the expectation of the latent variables. Provide a Bayesian interpretation for your answer.

$$E(m_{ic}|x,\lambda) = P(c|x,\lambda)\cdot 1$$

$$= P(c|x,\lambda)$$

$$= \frac{P(x|c,\lambda)\,P(c|\lambda)}{P(x|\lambda)} = \gamma_{ic}$$

$$\gamma_{ic} = \frac{\pi_k P(m_i;\lambda_k)}{\sum_{l=1}^{K} \pi_l P(m_i;\lambda_l)}$$

5 pts.

(c) **Bonus question:** Assume the expectations of the latent variables are given. Calculate the unknown parameters λ_c's. Write down the details of your calculations.

$$L(\theta, \pi | x) = \sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic} \log \pi_c P(x|\theta_c) - \lambda\left(\sum_c \pi_c - 1\right)$$

$$L = \sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic}\left[\log \pi_c + \log \frac{\lambda_c^{x_i} e^{-\lambda_c}}{x_i!}\right] - \lambda\left(\sum_c \pi_c - 1\right)$$

$x_i \log \lambda_c - \lambda_c$

5 pts.

$$\frac{\partial L}{\partial \lambda_c} = \sum_{i=1}^{n} y_{ic} \cdot \frac{x_i \lambda_c^{x_i-1} e^{-\lambda_c} + (-e^{-\lambda_c}) \lambda_c^{x_i}}{\left(\frac{\lambda_c^{x_i} e^{-\lambda_c}}{x_i!}\right)} = 0$$

$$\sum_{i=1}^{n} y_{ic}(x_i!)\left[\frac{x_i}{\lambda_c} - 1\right) = 0$$

$$\sum_{i=1}^{n} \frac{y_{ic} x_i}{\lambda_c} = \sum_{i=1}^{n} y_{ic}$$

$$\lambda_c = \frac{\sum y_{ic} x_i}{\sum y_{ic}}$$