

Introduction to Learning and Intelligent Systems

Final Exam

Aug 11, 2015

Time limit: 120 minutes

Number of pages: 20

Total points: 100

You can use the back of the pages if you run out of space. Collaboration on the exam is strictly forbidden. Please show *all* of your work and always *justify* your answers.

(1 point) Please fill in your name and student ID.

Please leave the table below empty.

Problem	Maximum points	Obtained
1.	15	
2.	15	
3.	9	
4.	10	
5.	15	
6.	20	
7.	15	
Total	100	

1. MAP Inference for Classification

(15 points)

In this problem we will work with the following probabilistic model for classification. If the features of the data point are $\mathbf{x} \in \mathbb{R}^d$, then the probability of class $y \in \{-1, +1\}$ is equal to

$$P(y | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y\mathbf{w}^T \mathbf{x})},$$

where $\mathbf{w} \in \mathbb{R}^d$ is the parameter vector. We will additionally use the following prior on \mathbf{w}

$$P(\mathbf{w}; b) = \frac{1}{Z} \exp\left(-\frac{\sum_{i=1}^d w_i^4}{b}\right),$$

where $b > 0$ is some constant, Z ensures that $P(\mathbf{w}; b)$ is normalized, and w_i denotes the i^{th} coordinate of the parameter vector \mathbf{w} . For the following questions, we will assume that we are given the i.i.d. training samples

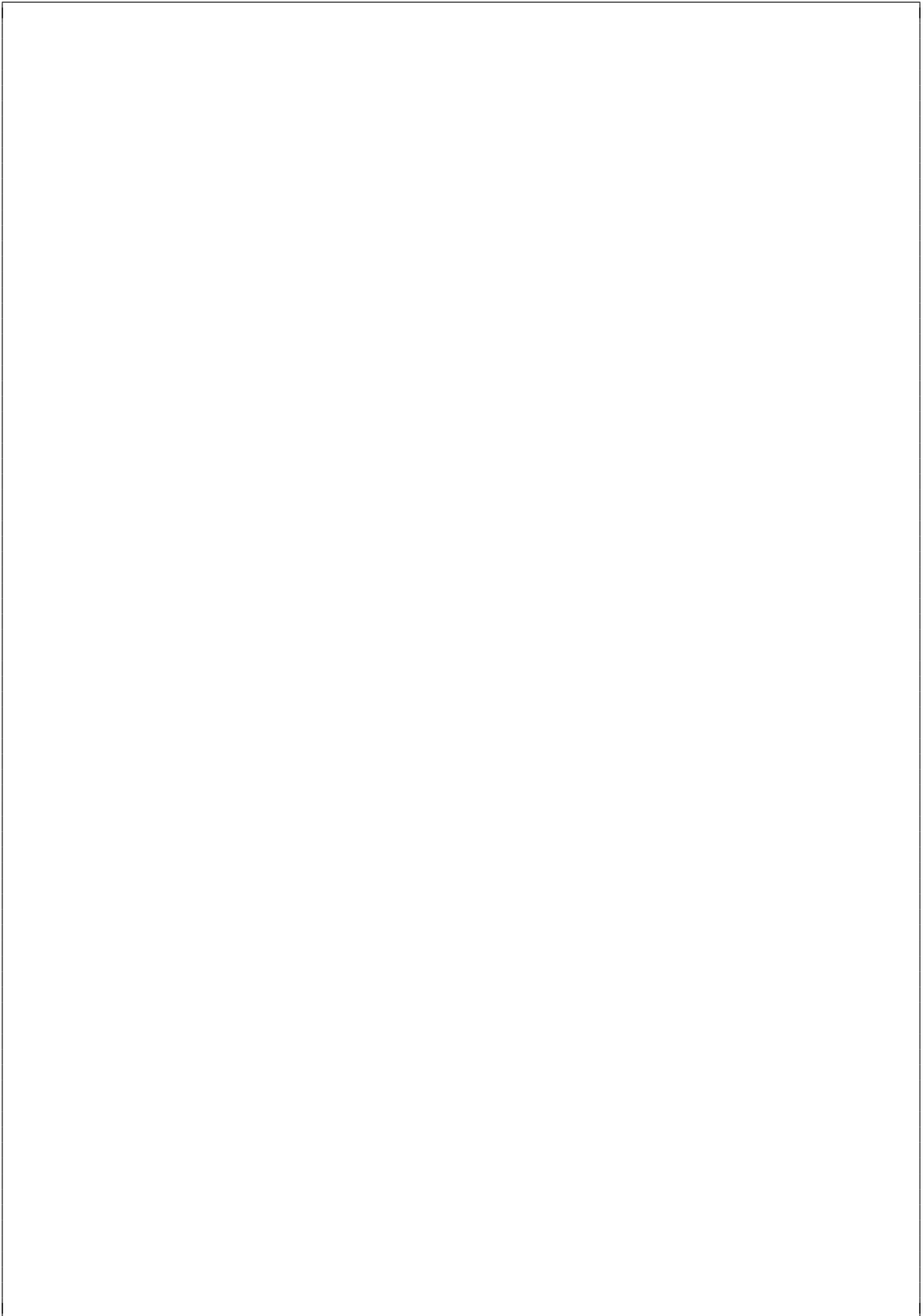
$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n).$$

- (5 points) (i) Show that MAP estimation, i.e. maximizing $\prod_{i=1}^n P(y_i | \mathbf{x}_i, \mathbf{w})P(\mathbf{w}; b)$ leads to the following regularized loss optimization problem

$$\underset{\mathbf{w}}{\text{minimize}} \underbrace{\sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))}_{\text{loss } L(\mathbf{w})} + \underbrace{\frac{1}{b} \sum_{i=1}^d w_i^4}_{\text{regularizer } C(\mathbf{w})}.$$

(4 points) (ii) Compute the gradient for the objective function $L(\mathbf{w}) + C(\mathbf{w})$ derived above.

(6 points) (iii) Write down the pseudo-code for stochastic gradient descent for finding the optimal parameters to the above optimization problem.



2. Naive Bayes Classifier

(15 points)

Suppose you want to classify tweets as either *relevant* or *irrelevant*. Every tweet is described by a bag of words \mathbf{x} given in vector form, i.e. $\mathbf{x} = (x_1, \dots, x_l)$, where $x_i \in \{0, 1\}$ indicates whether the i^{th} word is present in the tweet.

You are given a set D of n training examples, i.e.

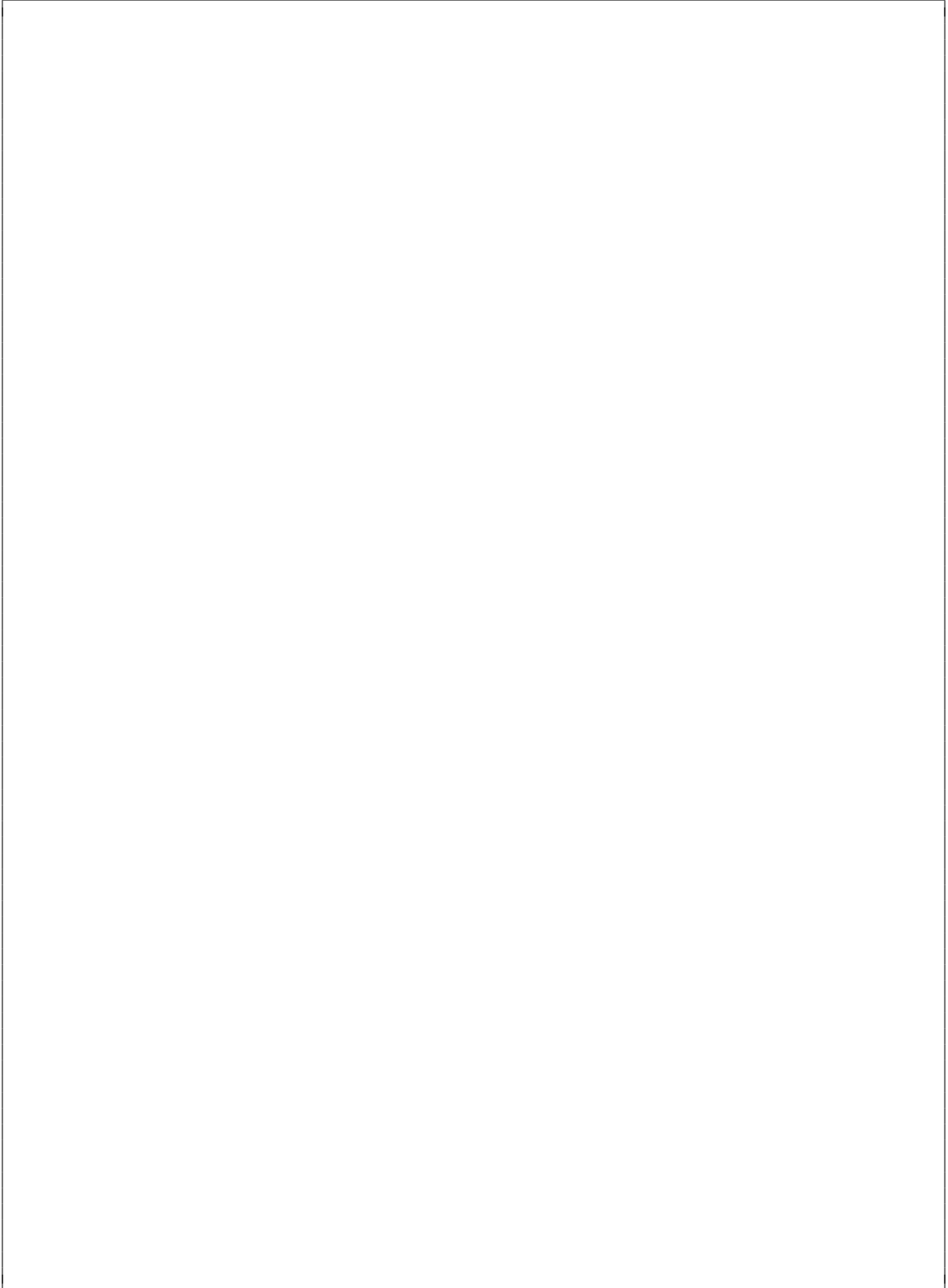
$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\},$$

where $\mathbf{x}^{(k)}$ is the bag of words for the k^{th} training sample and $y^{(k)} \in \{1, 0\}$ indicates whether the k^{th} training sample is relevant (1) or irrelevant (0).

- (4 points) (i) Write down the joint distribution $P(\mathbf{x}, y)$ of the naive Bayes classifier for the above setting. State the classification rule.

- (5 points) (ii) Assume that we have trained the model from the given data D using maximum likelihood estimation. State the resulting estimated prior $P(y)$ and the class-conditional distributions $P(x_i | y)$. No proofs are necessary, you can just state the resulting distributions.

(6 points) (iii) Consider the class posterior distribution $P(y | \mathbf{x})$ and assume that the cost $c_{1 \rightarrow 0}$ for classifying a relevant message as irrelevant is larger than the cost $c_{0 \rightarrow 1}$ of classifying an irrelevant message as relevant. The cost of classifying correctly is assumed to be zero. How does the classification rule change?



3. Kernels

(9 points)

Assume that for some space \mathcal{X} you are given $k_i : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, $i = 1, 2$. For each of the following statements either provide a proof that it is correct, or show a counterexample. (In your answers, you may directly use the kernel properties discussed in class.)

(3 points) (i) If k_1 and k_2 are kernels, then $k := k_1 + 2k_2$ is a kernel.

(3 points) (ii) If k_1 and k_2 are kernels, then $k := 3k_1 - k_2$ is a kernel.

(3 points) (iii) If k_1 and k_2 are kernels, and $\hat{k} := k_1 - 2k_2$ is a kernel, then $k := k_1 - k_2$ is also a kernel.

4. Artificial Neural Networks

(10 points)

Assume you want to classify the following four points $(x_1, x_2) \in \mathbb{R}^2$:

x_1	0	0	1	1
x_2	0	1	0	1
Class	+1	+1	+1	-1

(7 points) (i) In this question we will use activation units of the form

$$y = f_H(w_0 + \sum_{i=1}^n x_i w_i),$$

where f_H is the following threshold function

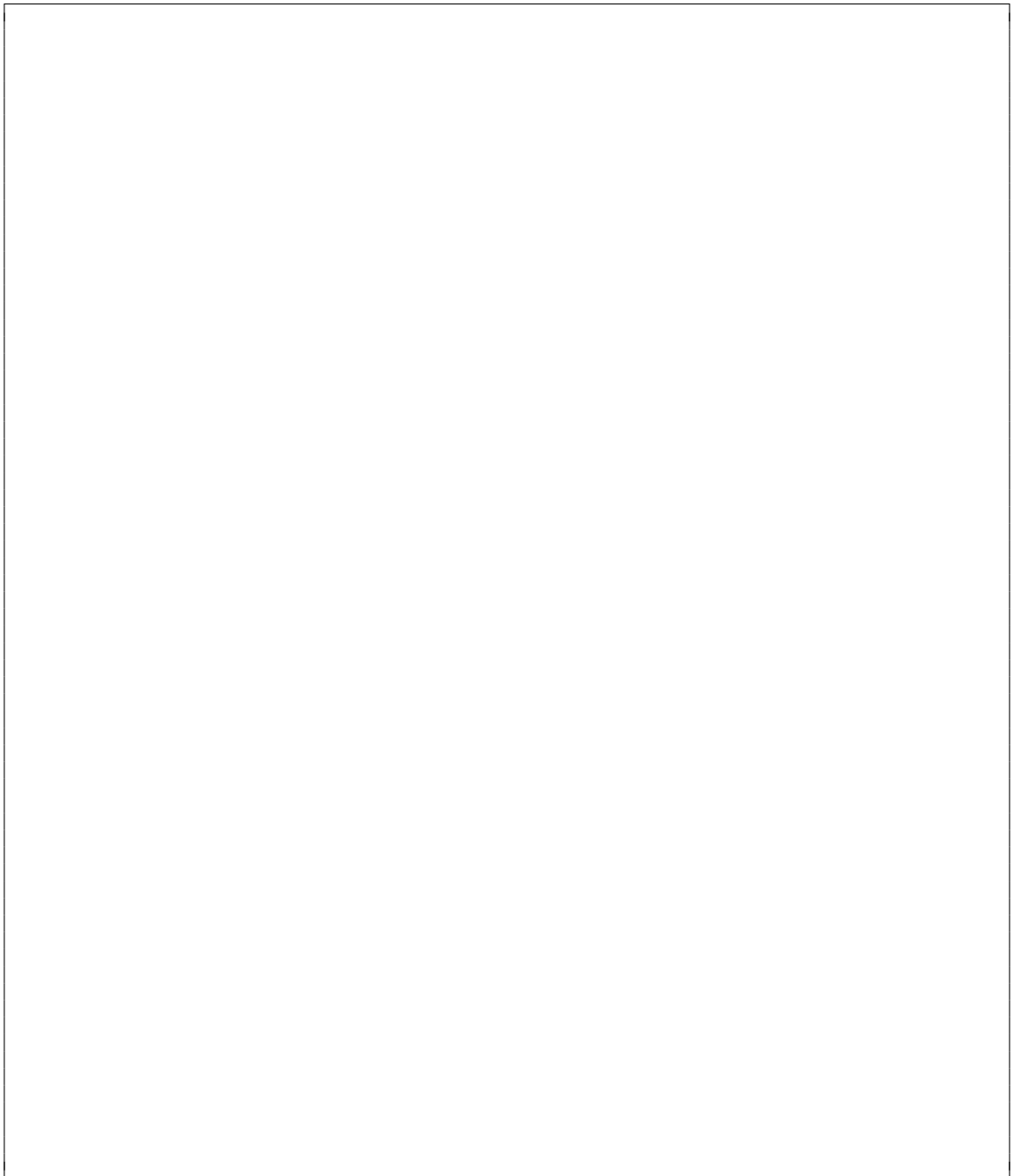
$$f_H(\alpha) = \begin{cases} -1 & \text{if } \alpha < 0 \\ +1 & \text{if } \alpha \geq 0 \end{cases}.$$

Remember that the numbers x_i are the inputs of the unit. Show by specifying a correct set of parameters that the above dataset can be perfectly classified with a single activation unit.

(3 points) (ii) Consider the following points:

x_1	0	0	1	1
x_2	0	1	0	1
Class	+1	-1	-1	+1

Provide a simple argument explaining why a neural network with one unit (of the type in (i)) cannot perfectly classify these points. (You can choose to draw a picture if you prefer).



5. Linear Regression

(15 points)

Remember that in linear regression our goal is to predict the outputs \mathbf{y} from the features X by minimizing $f(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$ with respect to the coefficients $\boldsymbol{\beta}$. The closed-form solution of the corresponding least-squares problem is given by


$$\boldsymbol{\beta} = (X^T X)^{-1} X^T \mathbf{y}. \quad (1)$$

For this problem, assume that we are given a single training point with one corresponding feature $x \in \mathbb{R}_{>0}$, and corresponding output $y \in \mathbb{R}$. Now, assume that we duplicate the given feature, so that our final problem has one training point with two (identical) features, and one output.

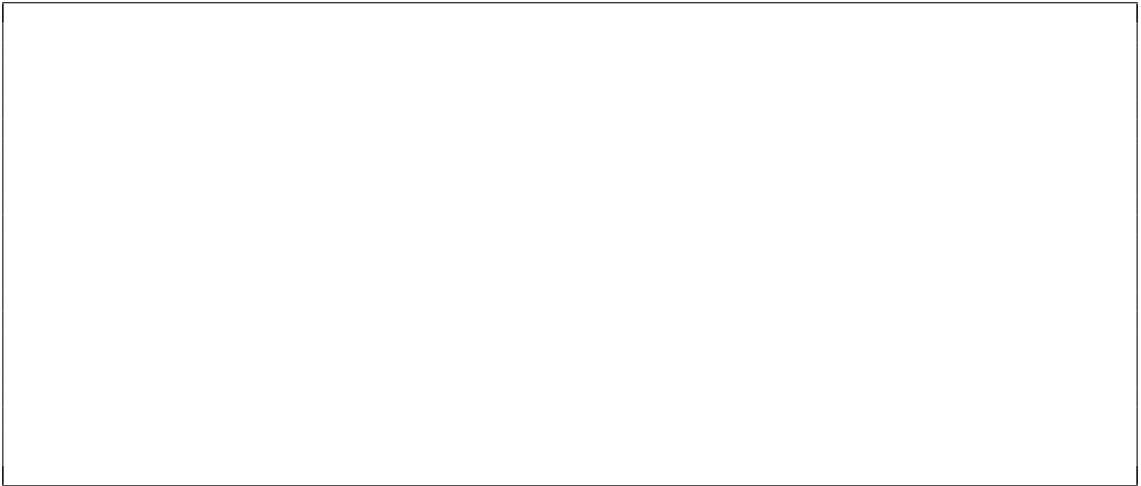
(2 points) (i) What are the dimensions (number of rows / columns) of X , \mathbf{y} , and $\boldsymbol{\beta}$ in this case?

(2 points) (ii) What is the rank of $X^T X$? Can we apply the closed form solution?

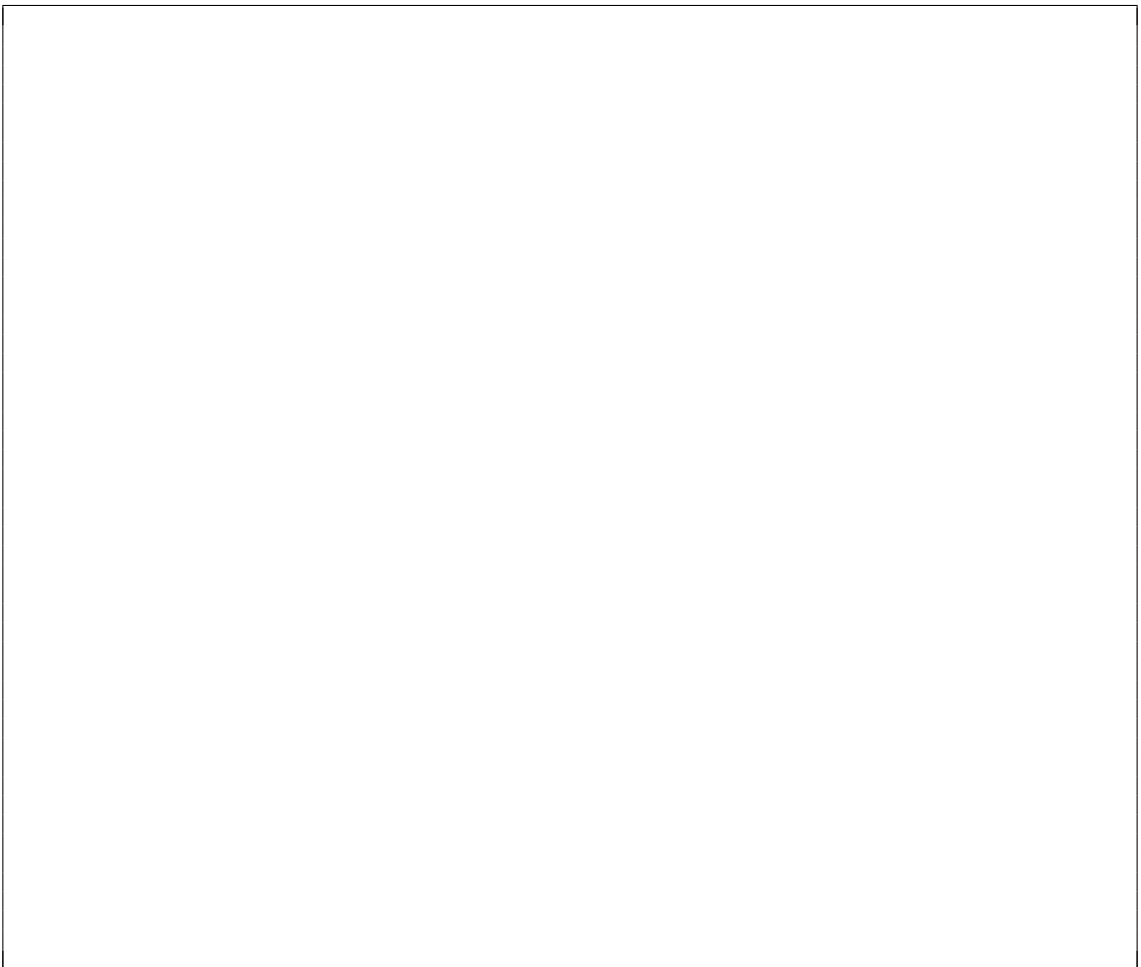
(5 points) (iii) Solve the linear regression problem for the set of data described in the introduction. Note that the objective $f(\beta)$ is convex, so that a point β^* is an optimum if and only if the gradient $\nabla f(\beta^*)$ is equal to zero. How many solutions do you get?



(3 points) (iv) Suppose we add an L_2 regularization term of the form $\lambda\|\beta\|_2^2$ for some $\lambda > 0$. Formulate the resulting optimization problem. How many solutions does this problem have?



(3 points) (v) Let β^* be a solution of the new problem. Describe the relationship that the components of β^* satisfy.

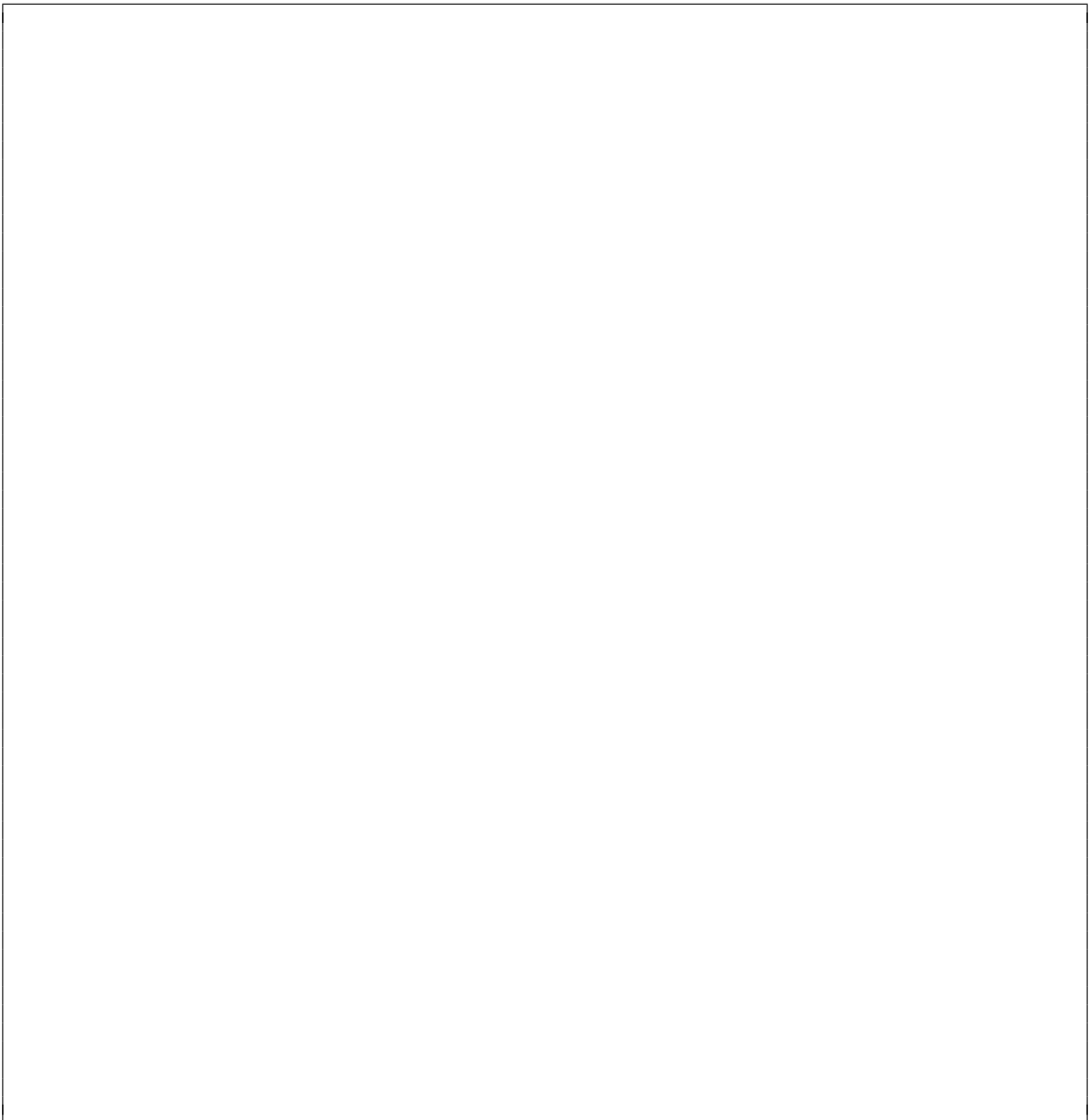


6. Markov chains and Hidden Markov Model

(20 points)

Harry the Naive is spending his well earned vacation camping T weeks in a swamp in Africa. Unfortunately, he did not consider the risk posed by mosquitoes: One out of five mosquitoes in the swamp carries a disease, which is contracted whenever the mosquito bites a visitor. Denote by $Y_t \in \{I, N\}$ the state of Harry in week $t = 1, 2, \dots, T$ where the state I signifies that Harry is infected with the disease and N that he is not. Initially, Harry arrives without the disease (i.e. $Y_1 = N$) but while he takes every precaution, at the end of every week, he is bitten by one random mosquito. If he is bitten at the end of week t and the disease is transmitted, then he will be infected in the subsequent week $t+1$ (i.e. $Y_{t+1} = I$) with no effect on his state in week t . Furthermore, Harry forgot to bring any medication which means that if he has contracted the disease, he will have it for the remainder of his stay.

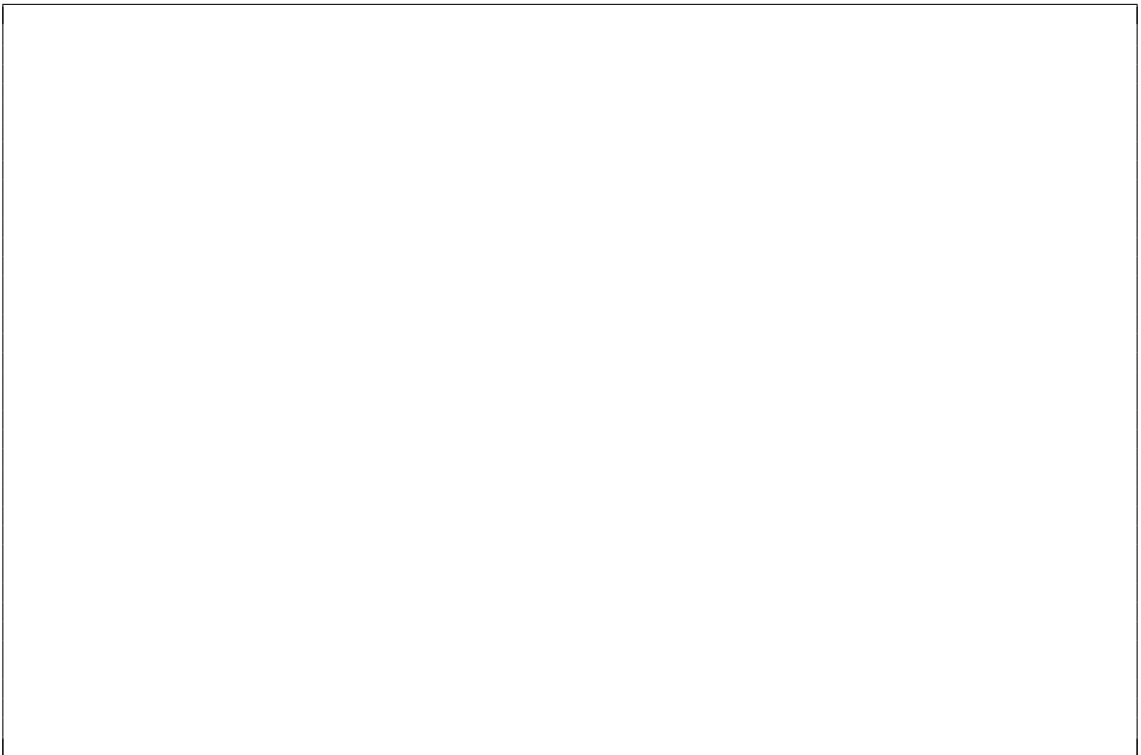
- (3 points) (i) Model the states Y_1, Y_2, \dots, Y_T as a Markov chain of order one describing Harry's vacation. In particular, specify $P(Y_1)$ as well as $P(Y_t | Y_{t-1})$ (for $t = 2, 3, \dots, T$) and draw the transition diagram of the Markov chain.



(2 points) (ii) What is the probability $P(Y_4 = N)$ that Harry is not infected in the fourth week?

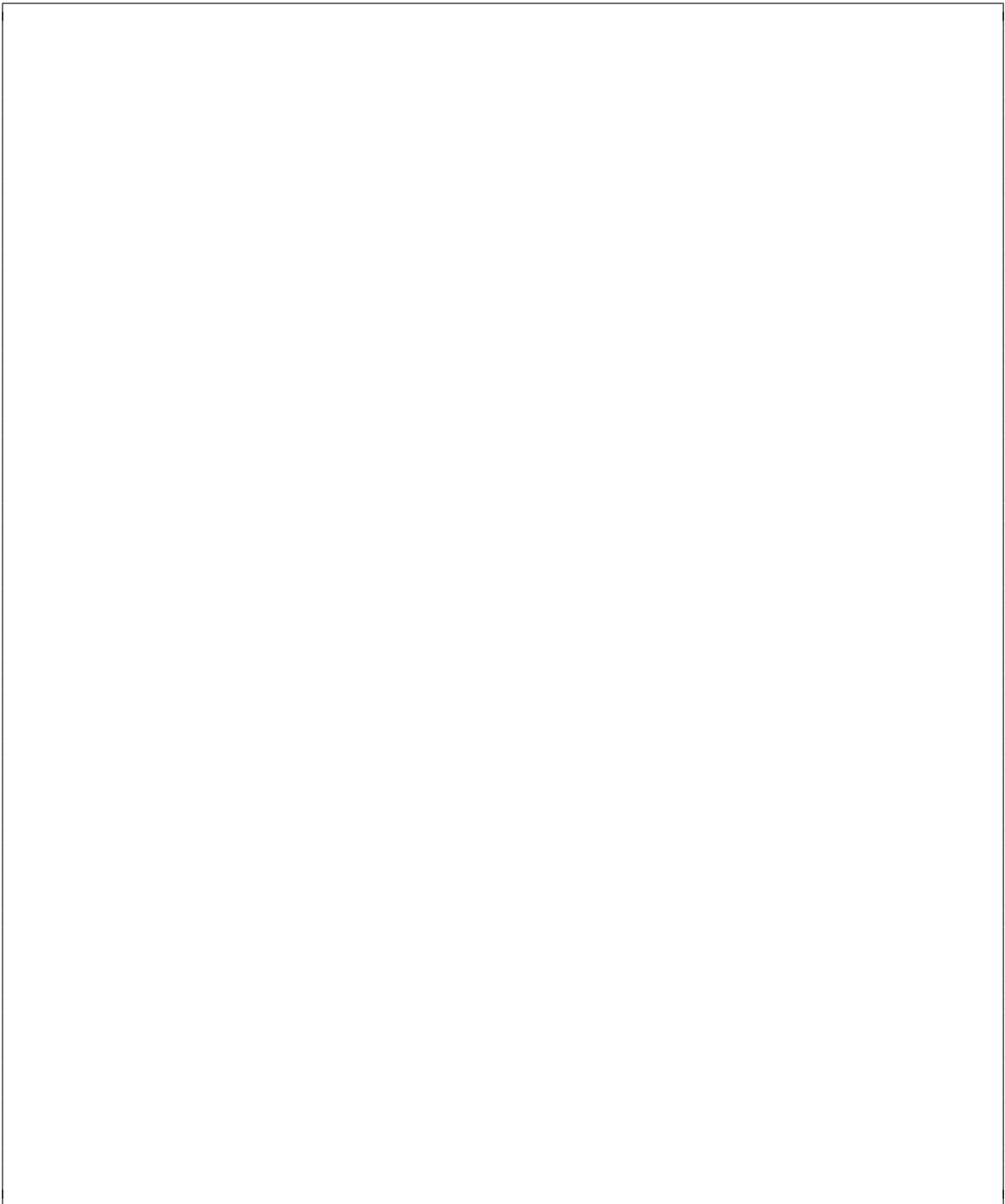


(2 points) (iii) Is the event that Harry is infected in the sixth week independent of whether he is infected in the fourth week given that he is not infected in the fifth week? Justify your answer.



Since Harry is no doctor, he cannot observe whether he is actually infected or not. Instead, he can only observe his well-being $X_t \in \{W, D, F\}$ in each week $t = 1, 2, \dots, T$. He can either feel well (i.e. $X_t = W$), dizzy (i.e. $X_t = D$) or exhibit fever (i.e. $X_t = F$). From Wikipedia, he knows that, if he is infected in a given week, he will have fever that week with probability 0.8 and otherwise feel dizzy. If not infected, he feels well with probability 0.5, dizzy with probability 0.4, and exhibits fever with probability 0.1.

(3 points) (iv) Formally specify Harry's stay as a *Hidden Markov Model*. What are the observed variables? What are the hidden variables? Write down $P(X_t | Y_t)$.



- (10 points)** (v) Suppose Harry goes on a four week trip which happens as follows: He has fever in the first week (i.e. $X_1 = F$) but then feels well in the second week (i.e. $X_2 = W$). In the third week, he feels dizzy (i.e. $X_3 = D$) and then has fever again in the fourth week (i.e. $X_4 = F$). At this point, Harry gets scared and flies back to Switzerland. His doctors in Switzerland now want to know what happened. What is the most likely sequence of states Y_1, Y_2, Y_3, Y_4 ?

7. Expectation Maximization

(15 points)

Kate is developing a new system and would like to understand the number of requests it has to serve. She counted the number of requests in $100ms$ intervals and observed that these counts follow a bimodal distribution. Hence, she thought about using a mixture model with *two* components. As her observations are positive (they are counts), she decided to model the components with Poisson distributions. Remember that a Poisson distribution X with parameter $\lambda > 0$ assigns the following probabilities

$$P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } \lambda \in \{0, 1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}.$$

In what follows, we will use the parameters from the following table.

Parameter	Description
$\pi \in [0, 1]$	The probability of the point being sampled from the first component
$\lambda_1 > 0$	The parameter of the first mixture
$\lambda_2 > 0$	The parameter of the second mixture

We will denote the observation by the random variable X , and the latent variable indicating the component the point is sampled from by Z , which takes on values in $\{1, 2\}$.

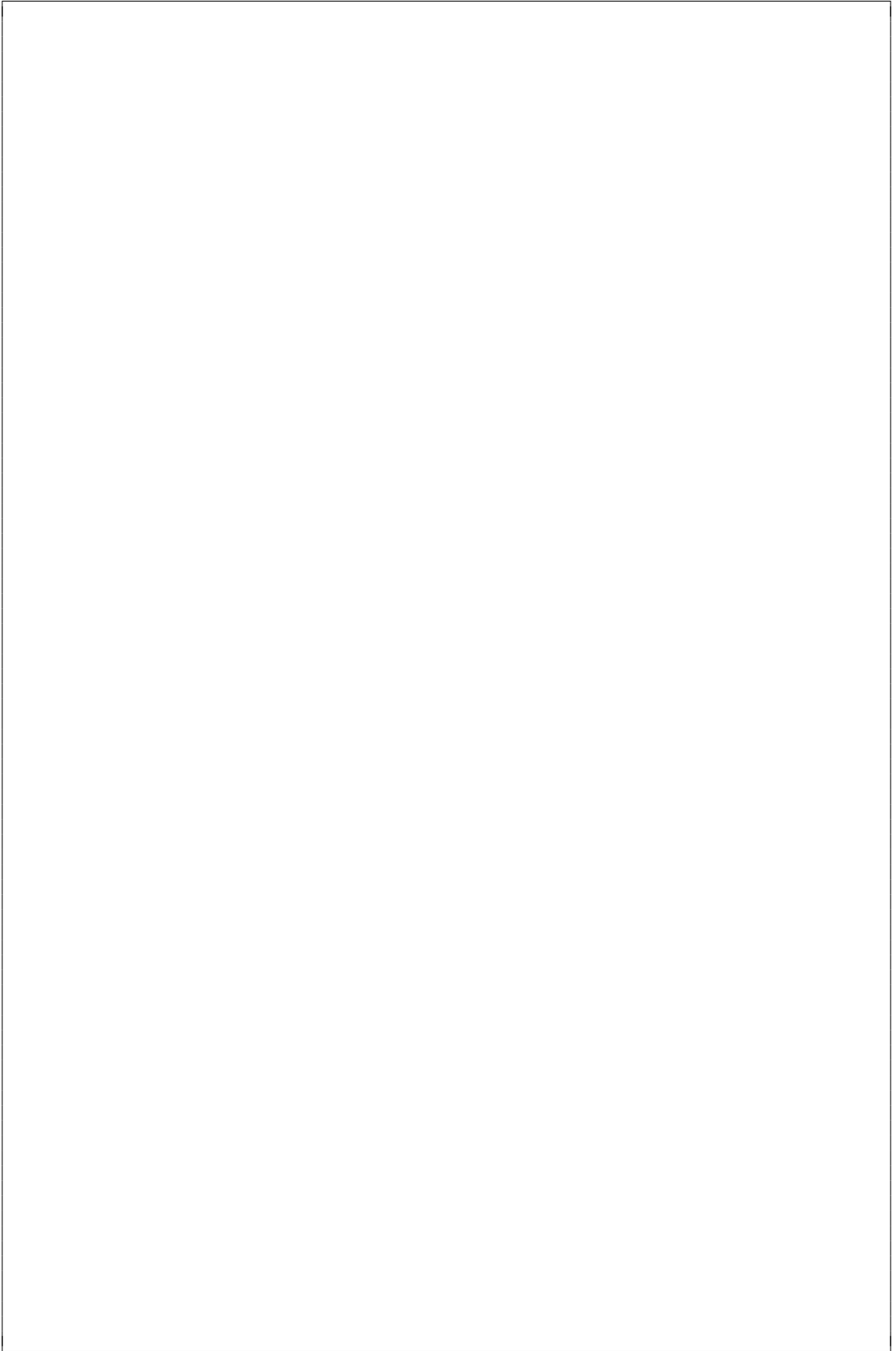
- (6 points) (i) After running several iterations of the EM algorithm, Kate obtained the following posterior probabilities in the E-step using the current set of parameters (please note that we denote the probabilities by $Q(\cdot)$).

Observation x_i	$Q(Z = 1 X = x_i)$	$Q(Z = 2 X = x_i)$
$x_1 = 2$	0.9	0.1
$x_2 = 4$	0.8	0.2
$x_3 = 8$	0.3	0.7

Write down the objective that the M-step is optimizing with respect to the parameters π , λ_1 and λ_2 . Remember that this is the expected log-likelihood of the complete data under the computed posterior distribution $Q(Z | X)$, or written formally

$$\mathbb{E}_{Q(Z|X)}[\log P(x_1, z_1, x_2, z_2, x_3, z_3)],$$

where we have denoted by $z_i \in \{1, 2\}$ the latent variable corresponding to the i -th data point, i.e. the component it has been sampled from.



(6 points) (ii) Perform the M-step update and compute the new parameters π , λ_1 and λ_2 .

(3 points) (iii) Kate would also like to experiment with different mixture components, but she does not know how to *compare* the resulting models. What strategy would you suggest to Kate? How should she choose a model? Please provide a principled approach.

