Last Name .... First Name......................

Exam for Pattern Classification and Machine Learning, 2012

Prof. Matthias Seeger

- You have 180 minutes in total
- Write your name in legible characters on the top of this page
- Write all your answers on the exam sheets (no extra sheets)
- No documentation allowed apart from 1 sheet A5 of your own notes
- No calculator (or any other electronic device) is allowed
- Have your student card displayed before you on your desk
- The exam has a total of 55 points
- The exam consists of 19 pages (10 sheets, double-sided). Please check that you have received all pages, numbered 1-19. Last two empty pages can be used for scratch notes

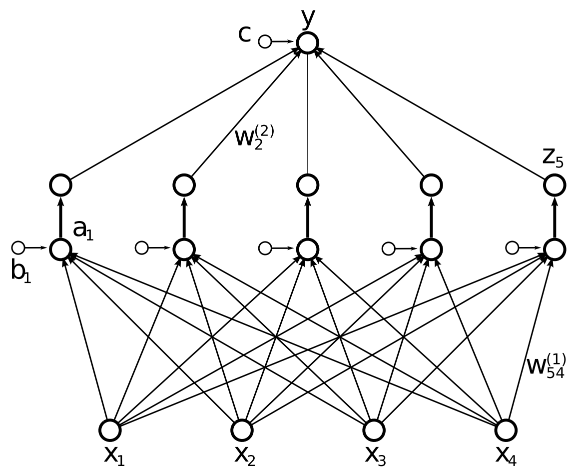
Distribution of points:

1. ..../7
2. ..../4
3. ..../6
4. ..../5
5. ..../6
6. ..../7
7. ..../6
8. ..../4
9. ..../10

Total: ..../55

1 Multilayer Perceptron (7pts)

Suppose you are given a training dataset $\{(\boldsymbol{x}_i, t_i) \,|\, i = 1, \ldots, n\}$, where the input vectors $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,4}]^T \in \mathbb{R}^4$, the targets $t_i \in \{-1, +1\}$. You use a multi-layer perceptron with one hidden layer of $h = 5$ units and transfer function $g(a) = \tanh(a)$, as depicted in the figure below.
Note: $x_1, \ldots, x_4$ in the figure are components of *one* input vector $\boldsymbol{x} = [x_j]$.



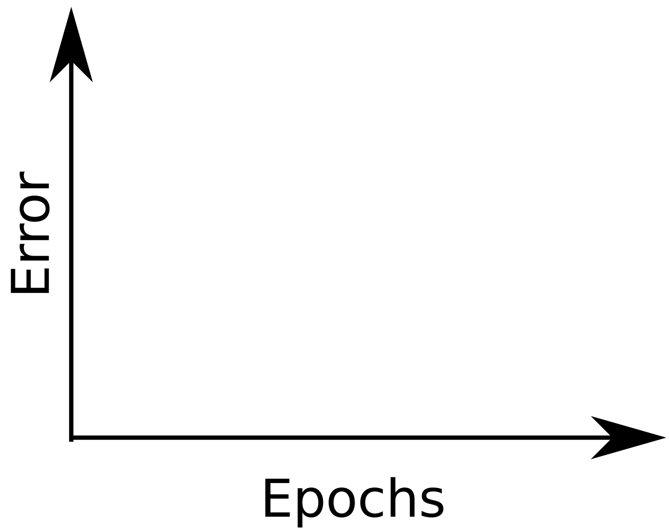
(a) Write down the forward equations $y = y(\boldsymbol{x}; \boldsymbol{w})$ for the model (here, $\boldsymbol{w}$ collects *all* parameters). This is easier if you use intermediate variables a_k and $z_k = g(a_k)$, $k = 1, \ldots, 5$. **(1pt)**

(b) What is the total number of parameters of this model? .... **(1pt)**

(c) Suppose you have $n = 200$ datapoints and use $h = 200$ hidden units. What problem are you likely to encounter when training the MLP?

....

How can you guard against this problem? Provide a brief explanation, and support your argument by drawing a qualitative example of the training set error and the validation set error in the figure below. **(2pts)**



(d) Your boss suggests to minimize the following error function:

$$\tilde{E} = \sum_{i=1}^{n} \frac{1}{5}(y(\boldsymbol{x}_i) - t_i)^5$$

How do you convince him that this is a bad idea? **(1pt)**

(e) You agree to use the following error function instead:

$$E = \sum_{i=1}^{n} E_i, \quad E_i = \frac{1}{6}(y(\boldsymbol{x}_i) - t_i)^6.$$

Compute the following gradient component for pattern i.
Hint: Weight $w_{54}^{(1)}$ links x_4 to a_5. **(2pts)**

$\frac{\partial E_i}{\partial w_{54}^{(1)}} =$ ....

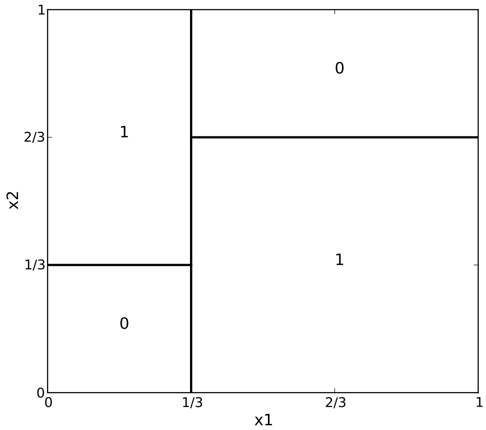
Space for Answer:

2 Decision Theory (4pts)

The random variables (x_1, x_2, t), $0 \leq x_1, x_2 \leq 1$, $t \in \{0, 1\}$ are distributed as follows. x_1, x_2 are independently and uniformly drawn from the interval $[0, 1]$. Given x_1, x_2, t is set as follows:

- If $x_1 \leq 1/3$: $\begin{cases} t = 1 & \text{if } x_2 > 1/3 \\ t = 0 & \text{if } x_2 \leq 1/3 \end{cases}$

- If $x_1 > 1/3$: $\begin{cases} t = 1 & \text{if } x_2 \leq 2/3 \\ t = 0 & \text{if } x_2 > 2/3 \end{cases}$



$(x_1, x_2) \mapsto t$ is visualized in the figure on the right.

(a) You observe x_2 only, but not x_1. Determine the conditional probability $P(t = 1|x_2)$. What is the Bayes optimal classifier $f^*(x_2) \to \{0, 1\}$? **(2pts)**

$P(t = 1|x_2) =$ ....

$$f^*(x_2) = \begin{cases} 0 & \text{if } \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 1 & \text{otherwise} \end{cases}$$

Space for Derivation:

(b) In the figure above (right), shade the set $\{(x_1, x_2)\}$ where the optimal classifier f^* commits an error. What is the Bayes error of f^*? **(2pts)**

Bayes error of f^*: ....

3 Perceptron (6pts)

We run the perceptron algorithm in 2D, to learn a linear discriminant

$$y(x_1, x_2) = w_1 x_1 + w_2 x_2 + w_3, \quad \boldsymbol{w} = [w_1, w_2, w_3]^T.$$

The bias term is w_3. The training dataset is $\{(\boldsymbol{x}_i, t_i) \mid i = 1, \ldots, n\}$, where $t_i \in \{-1, +1\}$, and $\boldsymbol{x}_i = [x_{i1}, x_{i2}, x_{i3}]^T$, where $x_{i3} = 1$ for all i (to accommodate w_3).

(a) Complete the missing definitions (.....) in the code below. **(1pt)**

---

repeat
  for $i \in \{1, \ldots, n\}$ (in random order) **do**

    if .... (HERE: Condition for update) **then**

      .... (HERE: Update of weights $\boldsymbol{w}$)

    end if
  end for

until .... (HERE: Condition for termination)

(b) Under which (necessary and sufficient) condition on the training set does the perceptron algorithm terminate after finitely many steps? **(1pt)**

Answer:

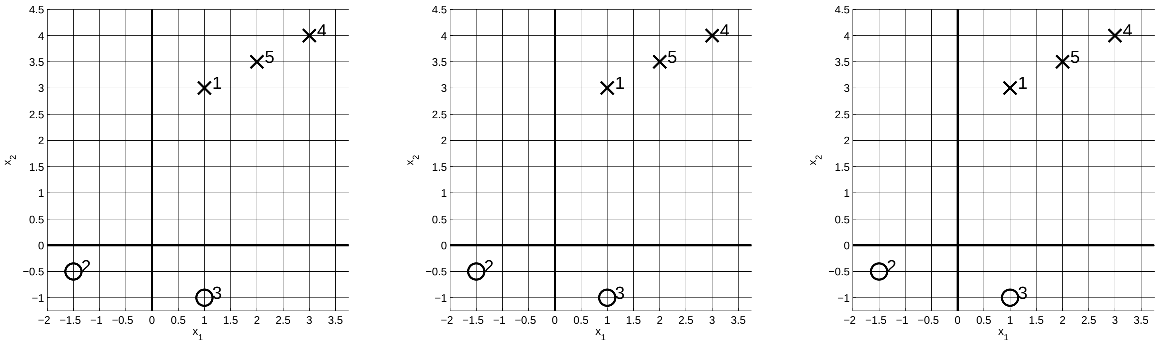
(c) You are given the following training dataset (also see figure on next page):

Order	x_{i1}	x_{i2}	x_{i3}	t_i
1	1	3	1	+1
2	$-\frac{3}{2}$	$-\frac{1}{2}$	1	-1
3	1	-1	1	-1
4	3	4	1	+1
5	2	$\frac{7}{2}$	1	+1

Run **three** steps of the perceptron algorithm, starting from the weight vector $\boldsymbol{w}_0 = [1, -1, 0]^T$, processing the datapoints $i = 1, 2, 3$ in this order.

Part (c) continues on next page

Part (c) continued



(A) After step 1 (B) After step 2 (C) After step 3

Report the weight vectors $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, $\boldsymbol{w}_3$ after each step:

After step 1: $\boldsymbol{w}_1 =$ ....

After step 2: $\boldsymbol{w}_2 =$ ....

After step 3: $\boldsymbol{w}_3 =$ ....

Also, **draw** the corresponding separating lines into the figures (A), (B), (C) above ($\times$ are $+1$, $\circ$ are -1). *Hint*: The line corresponding to $\boldsymbol{w} = [w_1, w_2, w_3]^T$ crosses the vertical axis at $x_2 = -w_3/w_2$. **(3pts)**

Space for calculations:

(d) Describe a preprocessing technique which can speed up the convergence of the perceptron algorithm in practice. Apply the technique to the 4th datapoint $\boldsymbol{x}_4 = [3, 4, 1]^T$. **(1pt)**

4th point after preprocessing: ....

Description:

4 Kernel Methods (5pts)

(a) Show that

$$K_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) = \left[\varepsilon^2 + \boldsymbol{x}^T \boldsymbol{y}\right]^2, \quad \varepsilon > 0,$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ are two-dimensional variables, is a valid kernel function that corresponds to a feature map $\boldsymbol{\phi}_\varepsilon(\boldsymbol{x}) \in \mathbb{R}^6$.

Derive the feature mapping $\boldsymbol{\phi}_\varepsilon(\boldsymbol{x})$. **(2pts)**

(b) A kernel method based on K_ε produces a function

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}_\varepsilon(\boldsymbol{x}), \qquad \boldsymbol{w} = [w_1, \ldots, w_6]^T.$$

Show how to derive a weight vector $\tilde{\boldsymbol{w}} = [\tilde{w}_1, \ldots, \tilde{w}_6]^T$ so that

$$\tilde{\boldsymbol{w}}^T \boldsymbol{\phi}_1(\boldsymbol{x}) = y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}_\varepsilon(\boldsymbol{x}),$$

meaning that $y(\boldsymbol{x})$ can be represented by the kernel K_1 (with $\varepsilon = 1$) as well. **(1pt)**

(c) A friend says: "From part (b), it follows that the spaces of functions $y(\boldsymbol{x})$ for kernel methods using K_ε are the same for all $\varepsilon > 0$. This means that running the SVM algorithm on some data will result in the same classifier, no matter what ε is." Explain the mistake in this argument. **(2pts)**

5 Naive Bayes Classification (6pts)

A binary Naive Bayes classifier (targets $t \in \{0, 1\}$, documents $\boldsymbol{x}$) uses seven binary features: $\boldsymbol{\phi}(\boldsymbol{x}) \in \{0, 1\}^7$. The training data is given by

$$\boldsymbol{\Phi}_0 = \overbrace{\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}}^{m=1,2,3,4,5,6,7}, \quad \boldsymbol{\Phi}_1 = \overbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}}^{m=1,2,3,4,5,6,7}.$$

There are 7 documents for class $t = 0$, 8 documents for class $t = 1$, the feature vectors are the rows of the data matrices ($\boldsymbol{\Phi}_0$ for class 0, $\boldsymbol{\Phi}_1$ for class 1).

(a) The Naive Bayes classifier has parameters $p_m^{(k)} = \Pr\{\phi_m(\boldsymbol{x}) = 1 | t = k\}$. Compute the maximum likelihood estimates $\hat{p}_m^{(k)}$, $m = 1, \dots, 7$, $k = 0, 1$. **(1pt)**

m	1	2	3	4	5	6	7
$\hat{p}_m^{(0)}$							
$\hat{p}_m^{(1)}$							

$\hat{P}(t = 0) =$ ....

$\hat{P}(t = 1) =$ ....

(b) A new document $\boldsymbol{x}_*$ gives rise to the feature vector $\boldsymbol{\phi}(\boldsymbol{x}_*) = [0, 0, 1, 1, 1, 0, 1]^T$. Using the trained Naive Bayes classifier, what is the probability that the document belongs to class 0? **(3pts)**

$$\hat{P}(t = 0 | \boldsymbol{x}_*) = \frac{1}{1 + a \cdot b^6}, \qquad a = \text{................}, \quad b = \text{................}$$

Space for calculations:

(c) Your data collection collaborator informs you about a bug in the data matrices $\mathbf{\Phi}_0$, $\mathbf{\Phi}_1$ above: for features $m = 2$ and $m = 3$, all values have been flipped ($0 \leftrightarrow 1$). Determine $\hat{P}(t = 0|\boldsymbol{x}_*)$ for the document $\boldsymbol{x}_*$ in (b) and the corrected data. **(2pts)**

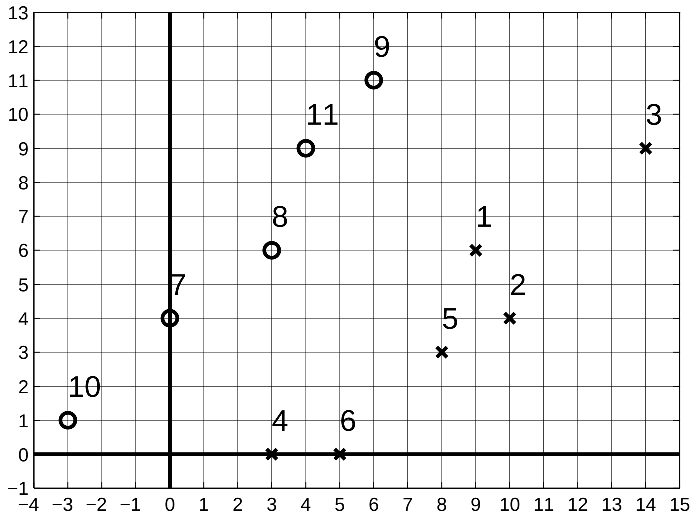
$$\hat{P}(t = 0|\boldsymbol{x}_*) = \frac{1}{1 + c \cdot d^6}, \qquad c = \ldots\ldots\ldots\ldots\ldots\ldots, \quad d = \ldots\ldots\ldots\ldots\ldots\ldots$$

Space for calculations:

6 Maximum Margin Perceptron and SVM (7pts)

Eleven data points representing two classes (crosses $t_i = +1$ and circles $t_i = -1$) are shown in the figure below. We use a maximum margin perceptron for classification:

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\boldsymbol{w}_0^T \boldsymbol{x} + b\right), \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$



(a) Draw the decision boundary of the maximum margin perceptron solution into the figure. **(1pt)**

(b) Provide the unit-norm weight vector $\boldsymbol{w}_0$, bias parameter b and margin κ of the optimal solution. **(3pts)**

$\boldsymbol{w}_0$ = ...

b = ...

κ = ...

Space for calculations

(c) How many support vectors are there in the optimal solution? Provide their indices. **(1pt)**

....

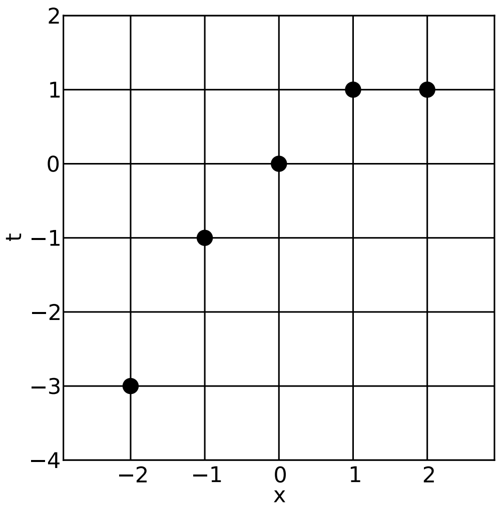
(d) Express the normalized weight vector $\boldsymbol{w}_0$ of (b) as linear combination of the support vectors found in (c). **(2pts)**
Hint: Consider the orthogonal projection of the class -1 support vector onto the line through the class $+1$ support vectors.

Result: $\boldsymbol{w}_0 =$ ....

Space for calculations

7 Least Squares Regression (6pts)

You are given the following dataset $\{(x_1, t_1), \ldots, (x_5, t_5)\}$, where $x_i, t_i \in \mathbb{R}$.



You choose a linear model

$$y(x) = ax + b.$$

To fit the model, you find $a, b \in \mathbb{R}$ which minimize the squared error function:

$$E(a, b) = \frac{1}{2} \sum_{i=1}^{n} (y(x_i) - t_i)^2.$$

(a) Begin by computing the following statistics from the data: **(1pt)**

$\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$\langle x^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$\langle xt \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i t_i \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$\langle t \rangle = \frac{1}{n} \sum_{i=1}^{n} t_i \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

Space for calculations:

(b) Derive the gradient components $\frac{\partial E}{\partial a}$ and $\frac{\partial E}{\partial b}$. Express them as functions of $\langle x \rangle$, $\langle x^2 \rangle$, $\langle xt \rangle$ and $\langle t \rangle$. **(2pts)**

$\frac{\partial E}{\partial a} \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$\frac{\partial E}{\partial b} \quad = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

Derivation (more space on following page):

Derivation for (b), continued:

(c) Compute the global minimum point (a_*, b_*) of the error $E(a, b)$. Draw the line corresponding to your solution into the figure on the previous page. **(2pts)**

$a_* =$ .... $b_* =$ ....

Calculation:

(d) Suppose you choose the model $y(x) = cx^2 + ax + b$ instead, with parameters $a, b, c \in \mathbb{R}$. Mark the correct answer from the following options, and **provide a brief explanation**. **(1pt)**

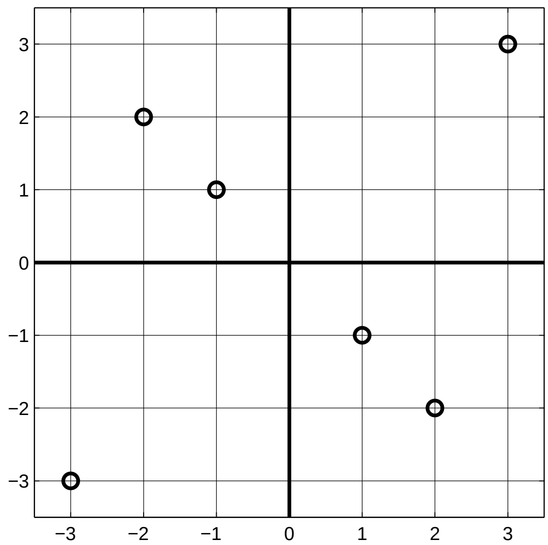
"The minimum squared error for the new setup, compared to the minimum error worked out above, ..."

( ) stays the same
( ) increases or stays the same
( ) decreases or stays the same
( ) can decrease or increase, depending on the data

Explanation:

8 Principal Components Analysis (4pts)

(a) How do you determine the *first* (leading) principal components direction $\boldsymbol{u}$ for data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$? Provide the definition of the matrix you use. **(1pt)**



(b) Consider the six datapoints in $\mathbb{R}^2$, depicted in the figure above. What is the *first* (leading) principal components direction $\boldsymbol{u}$? What is the corresponding eigenvalue λ? *Hint*: Consider the orthogonal directions $(1/\sqrt{2})[1, -1]^T$ and $(1/\sqrt{2})[1, 1]^T$. **(3pts)**

$\boldsymbol{u}$ = ...

λ = ...

Space for calculations on the following page

Space for calculations

9 Maximum Likelihood (10pts)

We model n positive data points $x_i > 0$ $(1 \leq i \leq n)$ as drawn independently from a probability distribution with density function

$$p(x|\gamma) = \frac{1}{2}\gamma^3 \, x^2 \, \exp(-\gamma \, x) \quad \text{for} \quad x > 0 \tag{1}$$

and $p(x|\gamma) = 0$ for $x \leq 0$. Here, $\gamma > 0$ is a parameter.

(a) Write down the likelihood function for the model (1) and the data. **(1pt)**

(b) Find the optimal value $\hat{\gamma}$ of γ by the principle of maximum likelihood. Write the result in the form **(2pts)**

$$\frac{1}{\hat{\gamma}} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Derivation:

(c) Write down a mixture model with four components $p(x|\omega_k)$, each of the form of Eq. 1, but with different γ_k. Denote the prior probabilities for the different components by $P(\omega_k)$, $k = 1, \ldots, 4$. **(1pt)**

$p_{\text{mixture}}(x) = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

(d) How many free and independent parameters does your mixture model of part (c) have? **(1pt)** ....

(e) For the mixture model in part (c), we have that $P(\omega_k) = 1/4$, $k = 1, \ldots, 4$, $\gamma_1 = 2$, and $\gamma_2 = \gamma_3 = \gamma_4 = 1$. Compute the posterior probability $P(\omega_1|x_i)$ of x_i coming from component 1, for $x_i = \log 2$ (natural logarithm: $e^{x_i} = 2$). **(2pts)**

$P(\omega_1|x_i = \log 2) =$ ....

Space for calculations:

(f) We want to estimate parameters by maximum likelihood. Derive an update equation for the parameter γ_1 of mixture component 1. To do so, define

$P(\omega_k|x_i) = \frac{p(x_i|\omega_k)P(\omega_k)}{p(x_i)}$ and show that for a constant C (provide it!) **(3pts)**

$$\frac{1}{\hat{\gamma}_1} = C\frac{\sum_i P(\omega_1|x_i)x_i}{\sum_i P(\omega_1|x_i)}$$

Note: Just determine the stationary point (no need for 2nd derivative).

Derivation:

Additional space for notes:

Additional space for notes: