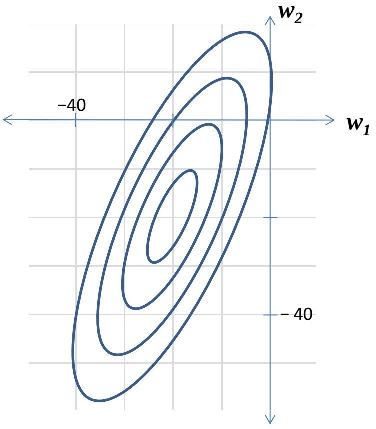
EE 769 Introduction to Machine Learning (IIT Bombay)

Mid-Semester Examination Solution Guide (Feb 29, 2020 8:30 – 10:30 am)

1. **L2 regularization:** Assume that the contours of an unregularized convex training loss function with respect to the parameters w_1 and w_2 are shown in the figure on the right.



   a. Trace the approximate locus of the optimal weights as the L2 regularization penalty on the weight vector $[w_1\ w_2]^T$ is increased from 0 to ∞, clearly indicating the least and the most regularized solutions. Answer by drawing on the figure in this sheet itself. [1]
   Join the center of the ellipses (zero regularization) with the origin (infinite regularization) (½ mark), and bend the curve up a bit (½ mark).

   b. Write the loss function for logistic regression with L2 regularization penalty on the weight vector assuming two weights, one bias, and N training samples (not related to this figure). [1]
   $L(\boldsymbol{w}) = \frac{C}{2}\|\boldsymbol{w}^T\boldsymbol{w}\|^2 - \sum_{n=1}^{N} t_n \log y_n + (1-t_n)\log(1-y_n)$ where
   $y_n = \frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x}-b)}$ ***(½ mark for regularization loss, ½ mark for cross entropy loss)***

   c. Given two regularization penalties $C_1$and C_2, how will you determine which is better in a general scenario? [1] ***Set aside some randomly selected training points for validation. Optimize the loss function over the remaining training points, but compare the performance on validation points for C=C$_1$ and C=C$_2$. (1 mark if validation is mentioned, ½ mark if testing is mentioned. The latter is wrong)***

2. **Bayesian classification:** Two class conditional densities are given by the following expressions:

$$p(x|c_0) = \begin{cases} k_0\sqrt{4-x^2}, \text{if} -2 < x < 2 \\ 0, \text{otherwise} \end{cases}, p(x|c_1) = \begin{cases} k_1\sqrt{9-(x-4)^2}, \text{if } 1 < x < 7 \\ 0, \text{otherwise} \end{cases}.$$

   a. What is the decision boundary if $(c_0) = \frac{4}{13}$? (Hint: Determine k_0/k_1 first, which is easy.) [2]
   At the decision boundary p(x|c$_0$)p(c$_0$) / p(x|c$_1$) / p(c$_1$) =1. Because probability densities should integrate to 1, k$_0$/k$_1$ = 9/4. Because priors add up to 1, p(c$_1$) = 9/13. Now, all that remains is $\sqrt{4-x^2} = \sqrt{9-(x-4)^2} \Rightarrow x^2 = (x-4)^2 - 5 = x^2 - 8x + 11 \Rightarrow x = 11/8$. ***(1 mark for k$_0$/k$_1$, ½ mark for p(c$_0$) and p(c$_1$), ½ mark for complete answer.)***

   b. For two multivariate Gaussian posterior densities given by $\frac{p(c_i)}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_i|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$, when is the decision boundary linear? [1] ***When ∑$_0$ = ∑$_1$. (binary marking)***

3. **Decision tree:** Examine the following set of 2-dimensional training samples and their class labels represented as a tuple $(x_{i,1}, x_{i,2}, t_i)$, where i is the sample index:
$(-2,2,0), (2,1,0), (7,-1,0), (9,-7,0), (0,-6,1), (5,-2,1), (8,-5,1)$. Now, answer the following:

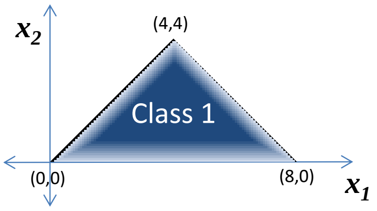
   a. What is the Gini impurity (hint: $p_1(1-p_1)$) of root/top/first node of a decision tree? [1]
   There are four points of one class, and 3 of another, so Gini is 3/7 x 4x7 = 12/49 = 0.245. (0 or 1 mark)

   b. What is the weighted average Gini impurity of children nodes after an optimal split of the root node? (Hint: try to draw the points in the answer sheet first.) [2]
   The optimal split has one pure node with three training points, and one node with 4 training points of which one is of a different class. So, weighted impurity is 3/7x0 + 4/7x(3/4x1/4) = 0.107 (1 mark for figuring out that one child will have zero impurity, and 1 mark for complete answer).

c. In the separate answer sheet, show the equivalence of Gini impurity for binary classification $p_1(1-p_1)$ and that for an arbitrary number of classes $1-\sum_c p_c^2$ (up to a constant), where c is class index. [2]
1 = (p₀+p₁)² = p₀²+p₁²+2p₀p₁ => 1 – p₀² – p₁² = 2p₁ p₀ =2 p₁ (1-p₁). Q.E.D. (0 or 2 marks)

4. **Feature normalization:** Assume a data matrix $\boldsymbol{X}$ for which $x_{i,j}$ is the j^{th} dimension of the i^{th} sample:
   a. Write the steps to obtain a data matrix $\widehat{\boldsymbol{X}}$ with mean and variance normalized. [1]
   Compute column-wise means μ_j and variances σ_j.
   Now, $x_{ij} \leftarrow (x_{ij} - \mu_j)/\sigma_j$ (0 or 1 mark)
   b. Is normalization necessary before training a linear classifier using L2 regularized logistic regression? Justify (hint: think of the metric being optimized). [1] ***Yes, because we take a dot product between weights and x in the loss term. (½ mark for correct answer, ½ mark for correct reason)***
   c. Is normalization necessary before training a decision tree using C4.5? Justify (hint: think of the metric being optimized). [1] ***No, because the decision threshold based on impurity relative to the order of training points sorted by any dimension does not change after a linear transformation of the data. (½ mark for correct answer, ½ mark for correct reason)***

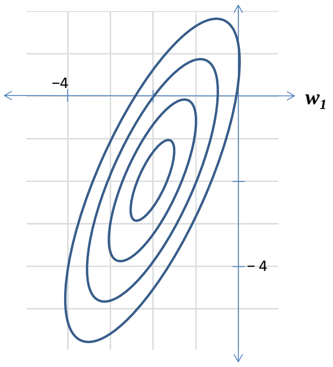
5. **Neural network:** Draw a neural network with a single hidden layer assuming sigmoid activation for all neurons, and indicate values of its weights and biases such that it can classify the given triangular region approximately as class 1, and the outside region as class 0. [3] ***Three hidden neurons will each have two weights (w,w), (w,0) and (w,-w), and appropriate biases to contain the triangle on one side. There will be one output node that will subject the addition of the three to a threshold close to 3. (½ mark for 2 inputs, 1 mark for 3 hidden nodes, 1 mark for correct weights of the hidden layer, ½ mark for getting the rest of it correct. If someone uses more than 3 hidden nodes, then the marking will be binary 0 or 3 marks.)***



6. **Kernelized SVM:** The expression for the output of a kernelized SVM is given by $f(\boldsymbol{x}) = b + \sum_{n=1}^{N} a_n\, t_n k(\boldsymbol{x}, \boldsymbol{x}_n)$, where N is the number of training points.
   a. Let N be 7, and let $b = 1, a_1 = 1, t_1 = 1, a_2 = 2, t_2 = 1, a_3 = \sqrt{2}, t_3 = -1$. Let the $a_n = 0$ for $n > 3$. How many support vectors does this SVM have? [1] ***Three (the first three training points). (0 or 1 mark)***
   b. Continuing with the details given in part (a), what is $f(\boldsymbol{x})$ for a test vector $\boldsymbol{x} = [2,2]^T$ further assuming that the training design (data) matrix is $\begin{bmatrix} 3 & -1 & 0 & 3 & 2 & -1 & 1 \\ 3 & 1 & 1 & 0 & 3 & -2 & 1 \end{bmatrix}^T$ and k is the cosine kernel (i.e., cos of the angle between the two inputs)? Is the point classified in the positive or negative class? [2]
   c. ***Take the cosine with the first three training points to get 1, 0, and 1/√2 respectively. Multiply by a_1t_1, a_2t_2, a_3t_3 respectively, then add b and examine the sign of f(x). (½ mark for correct answer, 1½ mark for correct f(x) computation).***

EE 769 Introduction to Machine Learning (IIT Bombay)

Mid-Semester Examination Solution Guide (Feb 29, 2020 8:30 – 10:30 am)



1. **L2 regularization:** Assume that the contours of an unregularized convex training loss function with respect to the parameters w_1 and w_2 are shown in the figure on the right.
   a. Trace the approximate locus of the optimal weights as the L2 regularization penalty on the weight vector $[w_1\ w_2]^T$ is increased from 0 to ∞, clearly indicating the least and the most regularized solutions. Answer by drawing on the figure in this sheet itself. [1] ***Join the center of the ellipses (zero regularization) with the origin (infinite regularization), and bend the curve up a bit. (½ mark), and bend the curve up a bit (½ mark).***

   b. Write the loss function for logistic regression with L2 regularization penalty on the weight vector assuming two weights, one bias, and N training samples (not related to this figure). [1]

   $L(\boldsymbol{w}) = \frac{C}{2}\|\boldsymbol{w}^T\boldsymbol{w}\|^2 - \sum_{n=1}^{N} t_n \log y_n + (1-t_n)\log(1-y_n)$ where

   $y_n = \frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x}-b)}$ ***(½ mark for regularization loss, ½ mark for cross entropy loss)***

   c. Given two regularization penalties $C_1$and C_2, how will you determine which is better in a general scenario? [1] ***Set aside some randomly selected training points for validation. Optimize the loss function over the remaining training points, but compare the performance on validation points for C=C$_1$ and C=C$_2$. (1 mark if validation is mentioned, ½ mark if testing is mentioned. The latter is wrong)***

2. **Bayesian classification:** Two class conditional densities are given by the following expressions:

   $p(x|c_0) = \begin{cases} k_0\sqrt{4-(x-4)^2}, \text{if } 2 < x < 6 \\ 0, \text{otherwise} \end{cases}, p(x|c_1) = \begin{cases} k_1\sqrt{9-x^2}, \text{if} -3 < x < 3 \\ 0, \text{otherwise} \end{cases}.$

   a. What is the decision boundary if $(c_0) = \frac{4}{13}$? (Hint: Determine k_0/k_1 first, which is easy.) [2] ***At the decision boundary p(x|c$_0$)p(c$_0$) / p(x|c$_1$) / p(c$_1$) =1. Because probability densities should integrate to 1, k$_0$/k$_1$ = 9/4. Because priors add up to 1, p(c$_1$) = 9/13. Now, all that remains is*** $\sqrt{4-(x-4)^2} = \sqrt{9-x^2}$ => $x^2 = (x-4)^2 + 5 = x^2 - 8x + 21$ => $= 21/8$ . . ***(1 mark for k$_0$/k$_1$, ½ mark for p(c$_0$) and p(c$_1$), ½ mark for complete answer.)***

   b. For two multivariate Gaussian posterior densities given by $\frac{p(c_i)}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_i|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$, when is the decision boundary linear? [1] ***When ∑$_0$ = ∑$_1$. (binary marking)***

3. **Decision tree:** Examine the following set of 2-dimensional training samples and their class labels represented as a tuple $(x_{i,1}, x_{i,2}, t_i)$, where i is the sample index: $(0,-6,0), (5,-2,0), (8,-5,0), (-2,2,1), (2,1,1), (7,-1,1), (9,-7,1)$. Now, answer the following:
   a. What is the Gini impurity (hint: $p_1(1-p_1)$) of root/top/first node of a decision tree? [1] ***There are four points of one class, and 3 of another, so Gini is 3/7 x 4x7 = 12/49 = 0.245. (binary marking)***
   b. What is the weighted average Gini impurity of children nodes after an optimal split of the root node? (Hint: try to draw the points in the answer sheet first.) [2] ***The optimal split has one pure node with three training points, and one node with 4 training points of which one is of a different class. So, weighted impurity is 3/7x0 + 4/7x(3/4x1/4) = 0.107 (1 mark for figuring out that one child will have zero impurity, and 1 mark for complete answer).***

c. In the separate answer sheet, show the equivalence of Gini impurity for binary classification $p_1(1-p_1)$ and that for an arbitrary number of classes $1-\sum_c p_c^2$ (up to a constant), where c is class index. [2] ***1 = $(p_0+p_1)^2 = p_0^2+p_1^2+2p_0p_1$ => $1 - p_0^2 - p_1^2 = 2p_1 p_0 = 2\, p_1\, (1-p_1)$. Q.E.D. (binary marking)***

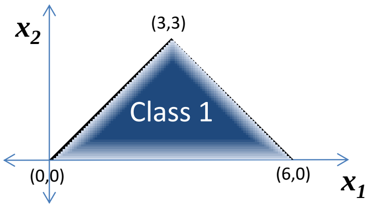
4. **Feature normalization:** Assume a data matrix $\boldsymbol{X}$ for which $x_{i,j}$ is the j^{th} dimension of the i^{th} sample:

   a. Write the steps to obtain a data matrix $\widehat{\boldsymbol{X}}$ with mean and variance normalized. [1]
   Compute column-wise means μ_j and variances σ_j.
   Now, $x_{ij} \leftarrow (x_{ij} - \mu_j)/\, \sigma_j$ (binary marking)

   b. Is normalization necessary before training a linear classifier using L2 regularized logistic regression? Justify (hint: think of the metric being optimized). [1]
   Yes, because we take a dot product between weights and x in the loss term. (½ mark for correct answer, ½ mark for correct reason)

   c. Is normalization necessary before training a decision tree using C4.5? Justify (hint: think of the metric being optimized). [1] ***No, because the decision threshold based on impurity relative to the order of training points sorted by any dimension does not change after a linear transformation of the data. (½ mark for correct answer, ½ mark for correct reason)***

5. **Neural network:** Draw a neural network with a single hidden layer assuming sigmoid activation for all neurons, and indicate values of its weights and biases such that it can classify the given triangular region approximately as class 1, and the outside region as class 0. [3] ***Three hidden neurons will each have two weights (w,w), (w,0) and (w,-w), and appropriate biases to contain the triangle on one side. There will be one output node that will subject the addition of the three to a threshold close to 3. (½ mark for 2 inputs, 1 mark for 3 hidden nodes, 1 mark for correct weights of the hidden layer, ½ mark for getting the rest of it correct. If someone uses more than 3 hidden nodes, then the marking will be binary 0 or 3 marks.)***



6. **Kernelized SVM:** The expression for the output of a kernelized SVM is given by $f(\boldsymbol{x}) = b + \sum_{n=1}^{N} a_n\, t_n k(\boldsymbol{x}, \boldsymbol{x}_n)$, where N is the number of training points.

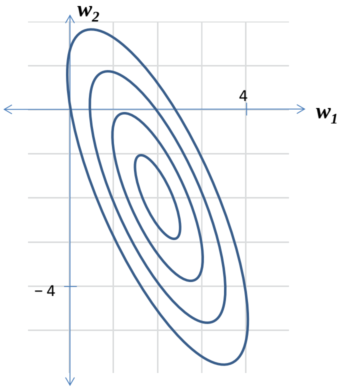
   a. Let N be 7, and let $b = 1, a_1 = 1, t_1 = 1, a_2 = 2, t_2 = 1, a_3 = \sqrt{2}, t_3 = -1$. Let the $a_n = 0$ for $n > 3$. How many support vectors does this SVM have? [1] ***Three (the first three training points). (binary marking)***

   b. Continuing with the details given in part (a), what is $f(\boldsymbol{x})$ for a test vector $\boldsymbol{x} = [2,2]^T$ further assuming that the training design (data) matrix is $\begin{bmatrix} 2 & -1 & 0 & 3 & 2 & -1 & 1 \\ 2 & 1 & 1 & 0 & 3 & -2 & 1 \end{bmatrix}^T$ and k is the cosine kernel (i.e., cos of the angle between the two inputs)? Is the point classified in the positive or negative class? [2] ***Take the cosine with the first three training points to get 1, 0, and 1/√2 respectively. Multiply by a_1t_1, a_2t_2, a_3t_3 respectively, then add b and examine the sign of f(x). (½ mark for correct answer, 1½ mark for correct f(x) computation).***

EE 769 Introduction to Machine Learning (IIT Bombay)

Mid-Semester Examination Solution Guide (Feb 29, 2020 8:30 – 10:30 am)

1. **L2 regularization:** Assume that the contours of an unregularized convex training loss function with respect to the parameters w_1 and w_2 are shown in the figure on the right.



   a. Trace the approximate locus of the optimal weights as the L2 regularization penalty on the weight vector $[w_1\ w_2]^T$ is increased from 0 to ∞, clearly indicating the least and the most regularized solutions. Answer by drawing on the figure in this sheet itself. [1] ***Join the center of the ellipses (zero regularization) with the origin (infinite regularization), and bend the curve up a bit (½ mark), and bend the curve up a bit (½ mark).***

   b. Write the loss function for logistic regression with L2 regularization penalty on the weight vector assuming two weights, one bias, and N training samples (not related to this figure). [1]

   $L(\boldsymbol{w}) = \frac{C}{2}\|\boldsymbol{w}^T\boldsymbol{w}\|^2 - \sum_{n=1}^{N} t_n \log y_n + (1-t_n)\log(1-y_n)$ where

   $y_n = \frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x}-b)}$ ***(½ mark for regularization loss, ½ mark for cross entropy loss)***

   c. Given two regularization penalties $C_1$and C_2, how will you determine which is better in a general scenario? [1] ***Set aside some randomly selected training points for validation. Optimize the loss function over the remaining training points, but compare the performance on validation points for C=C₁ and C=C₂. (1 mark if validation is mentioned, ½ mark if testing is mentioned. The latter is wrong)***

2. **Bayesian classification:** Two class conditional densities are given by the following expressions:

$$p(x|c_0) = \begin{cases} k_0\sqrt{4-(x-3)^2}, \text{if } 1 < x < 5 \\ 0, \text{otherwise} \end{cases}, p(x|c_1) = \begin{cases} k_1\sqrt{9-(x+1)^2}, \text{if} -4 < x < 2 \\ 0, \text{otherwise} \end{cases}.$$

   a. What is the decision boundary if $(c_1) = \frac{9}{13}$? (Hint: Determine k_0/k_1 first, which is easy.) [2] ***At the decision boundary p(x|c₀)p(c₀) / p(x|c₁) / p(c₁) =1. Because probability densities should integrate to 1, k₀/k₁ = 9/4. Because priors add up to 1, p(c₁) = 9/13. Now, all that remains is*** $\sqrt{4-(x-3)^2} = \sqrt{9-(x+1)^2}$ => $x^2-6x+9-4 = x^2+2x+1-9$ => $= 13/8$. . ***(1 mark for k₀/k₁, ½ mark for p(c₀) and p(c₁), ½ mark for complete answer.)***

   b. For two multivariate Gaussian posterior densities given by $\frac{p(c_i)}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_i|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T{\boldsymbol{\Sigma}_i}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$, when is the decision boundary linear? [1] ***When ∑₀ = ∑₁. (binary marking)***

3. **Decision tree:** Examine the following set of 2-dimensional training samples and their class labels represented as a tuple $(x_{i,1}, x_{i,2}, t_i)$, where i is the sample index: $(-2,2,1), (2,1,1), (7,-1,1), (9,-7,1), (0,-6,0), (5,-2,0), (8,-5,0)$. Now, answer the following:

   a. What is the Gini impurity (hint: $p_1(1-p_1)$) of root/top/first node of a decision tree? [1] ***There are four points of one class, and 3 of another, so Gini is 3/7 x 4x7 = 12/49 = 0.245. (binary marking)***

   b. What is the weighted average Gini impurity of children nodes after an optimal split of the root node? (Hint: try to draw the points in the answer sheet first.) [2] ***The optimal split has one pure node with three training points, and one node with 4 training points of which one is of a different class. So, weighted impurity is 3/7x0 + 4/7x(3/4x1/4) = 0.107 (1 mark for figuring out that one child will have zero impurity, and 1 mark for complete answer).***

c. In the separate answer sheet, show the equivalence of Gini impurity for binary classification $p_1(1-p_1)$ and that for an arbitrary number of classes $1-\sum_c p_c^2$ (up to a constant), where c is class index. [2] ***$1 = (p_0+p_1)^2 = p_0^2+p_1^2+2p_0p_1$ => $1 - p_0^2 - p_1^2 = 2p_1 p_0 = 2 p_1 (1-p_1)$. Q.E.D. (binary marking)***

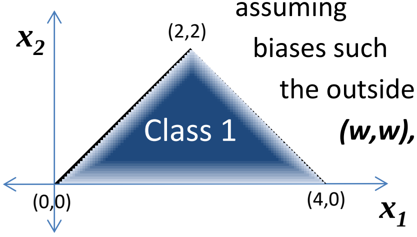
4. **Feature normalization:** Assume a data matrix $\boldsymbol{X}$ for which $x_{i,j}$ is the j^{th} dimension of the i^{th} sample:

a. Write the steps to obtain a data matrix $\widehat{\boldsymbol{X}}$ with mean and variance normalized. [1]
Compute column-wise means μ_j and variances σ_j.
Now, $x_{ij} \leftarrow (x_{ij} - \mu_j)/ \sigma_j$ (binary marking)

b. Is normalization necessary before training a decision tree using C4.5? Justify (hint: think of the metric being optimized). [1] ***No, because the decision threshold based on impurity relative to the order of training points sorted by any dimension does not change after a linear transformation of the data. (½ mark for correct answer, ½ mark for correct reason)***

c. Is normalization necessary before training a linear classifier using L2 regularized logistic regression? Justify (hint: think of the metric being optimized). [1] ***Yes, because we take a dot product between weights and x in the loss term. (½ mark for correct answer, ½ mark for correct reason)***

5. **Neural network:** Draw a neural network with a single hidden layer assuming sigmoid activation for all neurons, and indicate values of its weights and biases such that it can classify the given triangular region approximately as class 1, and the outside region as class 0. [3] ***Three hidden neurons will each have two weights (w,w), (w,0) and (w,-w), and appropriate biases to contain the triangle on one side. There will be one output node that will subject the addition of the three to a threshold close to 3. (½ mark for 2 inputs, 1 mark for 3 hidden nodes, 1 mark for correct weights of the hidden layer, ½ mark for getting the rest of it correct. If someone uses more than 3 hidden nodes, then the marking will be binary 0 or 3 marks.)***



6. **Kernelized SVM:** The expression for the output of a kernelized SVM is given by $f(\boldsymbol{x}) = b + \sum_{n=1}^{N} a_n\, t_n k(\boldsymbol{x}, \boldsymbol{x}_n)$, where N is the number of training points.

a. Let N be 7, and let $b = -1$, $a_1 = 1, t_1 = 1, a_2 = 2, t_2 = 1, a_3 = \sqrt{2}, t_3 = -1$. Let the $a_n = 0$ for $n > 3$. How many support vectors does this SVM have? [1] ***Three (the first three training points). (binary marking)***

b. Continuing with the details given in part (a), what is $f(\boldsymbol{x})$ for a test vector $\boldsymbol{x} = [2,2]^T$ further assuming that the training design (data) matrix is $\begin{bmatrix} 3 & -1 & 0 & 3 & 2 & -1 & 1 \\ 3 & 1 & 1 & 0 & 3 & -2 & 1 \end{bmatrix}^T$ and k is the cosine kernel (i.e., cos of the angle between the two inputs)? Is the point classified in the positive or negative class? [2] ***Take the cosine with the first three training points to get 1, 0, and 1/√2 respectively. Multiply by a_1t_1, a_2t_2, a_3t_3 respectively, then add b and examine the sign of f(x). (½ mark for correct answer, 1½ mark for correct f(x) computation).***