6.867: Exam 1, Fall 2017

Solutions

These are not the **only** acceptable answers. Some other answers also received credit.

Answer the questions in the spaces provided. Show your work neatly. **We will only grade answers that appear in the answer boxes or on answer lines.**

If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.

You may prepare and use both sides of one 8.5 inch x 11 inch sheet of paper upon which you may write/print anything you like. You may not use any electronic device or any other resource other than your two-sided sheet of paper.

Write your name on every page.

Come to the front if you need to ask a question.

Name: _____ MIT Email: _____

Question	Points	Score
1	32	
2	34	
3	34	
Total:	100	

Name: _____

Multiple Choice

1. (32 points) The first part of this exam consists of eight multiple choice questions. Each is worth 4 points. Answers preceded by a circle imply *only ONE* answer should be selected. Squares are for multi-choice questions and imply that you should select *ALL* valid answers.

   (a) Suppose we are interested in performing ridge regression, but with an unpenalized intercept term. That is, we want to minimize the following objective function:

   $$\sum_{i=1}^{n}(y_i - w_0 - (w_1, ..., w_p)^T \phi(x_i))^2 + \lambda \|(w_1, ..., w_p)\|^2 \quad (1)$$

   For sufficiently large λ, what will your estimate $\hat{y}_i$ be for y_i?

   - ◯ 0
   - ◯ $\frac{1}{n}\sum_{j=1}^{n} x_j$
   - ◯ $\frac{1}{n}\sum_{j=1}^{n} y_j$
   - ◯ The ith component of the projection of $(y_1, ..., y_n)$ onto the column space of $[\phi(x_1)...\phi(x_n)]$
   - ◯ Depends on i but none of the above

   Explanation: _____

   Solution: Third choice.
   For large λ, the $w_1, ..., w_p \to 0$. So we are approximately solving $\sum_{i=1}^{n}(y_i - w_0)^2$
   And the minimizing w_0 is $\frac{1}{n}\sum_{j=1}^{n} y_j$.
   And our estimator is $\hat{y}_i(x) = w_0$

   (b) Late at night, student Rick comes up with an idea for a new kernel: Let s_i be the number of hours student i slept last night. Then, define the kernel as $K(s_i, s_j) := s_i - s_j$

   Is this a valid kernel?

   - ◯ Valid kernel. Explanation: _____
   - ◯ Invalid kernel. Explanation: _____

   Solution: Invalid. It's not symmetric.

   (c) Rick is now trying to do classification: $y_i \in \{0, 1\}$ and $x_i \in \mathbb{R}$

   But he slept through all of the classification lectures, so he decides to solve classification using regression. That is, he ignores the fact that y_i is binary, and fits a linear regression function via least squares. The resulting regression function is:

   $\hat{y} = f(x; w) = w_0 + xw_1$

   Rick uses the decision rule: label 1 if $f(x; w) > 1/2$; and label 0 otherwise

   Suppose the training data is linearly separable. Is Rick's decision rule (with associated regression function) guaranteed to classify the training data without error?

◯ Yes. Provide a short argument:

◯ No. Provide a counterxample:

A short argument for YES or a counterexample for NO:

Solution: No

Consider the dataset:
$(-1/2, 0)$ repeated n times
(0,1)
(1,1) repeated n times

The data is linearly separable in X (for example, by $X = \frac{-1}{4}$).
As $n \to \infty$, the best fit line $\to \frac{2x}{3} + \frac{1}{3}$

Thus, for large n, the prediction at x = 0 will be class 0 since $\frac{2*0}{3} + \frac{1}{3} = 1/3 < 1/2$
So the middle point (which is of class 1) will be misclassified by Rick's decision rule.

(d) Assume we have N data points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ and M labels: $y_i \in 1, 2, \ldots, M$. We would like to extend binary-class perceptron to a multi-class perceptron. Below we provide one possible algorithm for multi-class perceptron which we run for T iterations.

1: Initialize $\boldsymbol{w}_y = 0$ for $y \in \{1, \ldots, M\}$
2: **for** $t = 1$ to T **do**
3: **for** $i = 1$ to N **do**
4: $z = \text{argmax}_y \langle \boldsymbol{w}_y, \boldsymbol{x}_i \rangle$
5: **if** $z \neq y_i$ **then**
6: $\boldsymbol{w}_{y_i} = \boldsymbol{w}_{y_i} - \boldsymbol{x}_i$
7: $\boldsymbol{w}_z = \boldsymbol{w}_z + \boldsymbol{x}_i$

Choose the lines (if any) of the algorithm which are incorrect:

- ☐ Line 4
- ☐ Line 6
- ☐ Line 7
- ☐ The algorithm is correct

Solution: If we assume line 4 is correct then lines 6 and 7 are incorrect. The correct solution is $\boldsymbol{w}_{y_i} = \boldsymbol{w}_{y_i} + \boldsymbol{x}_i$ and $\boldsymbol{w}_z = \boldsymbol{w}_z - \boldsymbol{x}_i$. That is because we need to lower score of wrong answer and raise the score of right answer. However, if we assume line 4 is incorrect then lines 6 and 7 will be correct. We accepted both answers (line 4 is incorrect or lines 6 and 7 are incorrect) in the exam.

(e) Assume we have a linear classification problem for which we run the following algorithms on the training dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. We obtain a set of weights $\boldsymbol{w} = (w_1, \ldots, w_p)$ and a bias term w_0. Next, we transform each feature j for all data points i by setting $x_i'^{(j)} = \frac{x_i^{(j)} - a}{b}$ where $a > 0$ and $b > 0$. We rerun the algorithms with the new dataset and obtain the new optimal feature weights w'.

In which of the algorithms the equation $w_j' = b \cdot w_j$ (for $j \neq 0$) **does not** hold. Please select all that are correct:

- ☐ Linear regression (no regularization)
- ☐ Ridge regression
- ☐ Logistic regression (no regularization)
- ☐ Hard-margin SVM

Does this transformation change the linear separability of the data?

- ◯ Yes. Explanation: _____
- ◯ No. Explanation: _____

Solution: Equation $w_j' = b \cdot w_j$ (for $j \neq 0$) does not hold for ridge regression.

For linear regression without regularization, we have the following loss function:

$$\sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2 = \sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^T\mathbf{a} + b\mathbf{w}^T(\frac{\mathbf{x_i} - \mathbf{a}}{b})))^2,$$

which shows subtracting a constant from $\mathbf{x}_i$ only changes the intercept and dividing $\mathbf{x}_i$ by b makes $w'_j = b \cdot w_j$ (for $j \neq 0$).
However, for the ridge regression with parameter λ and in the simplified case where $a = 0$, we have the optimal weights for the transformed dataset:

$$\mathbf{w}' = (X'^T X' + \lambda \mathbf{I})^{-1} X'^T Y \neq b(X^T X + \lambda \mathbf{I})^{-1} X^T Y = b \cdot \mathbf{w}.$$

For logistic regression without regularization, we have the following loss (negative log-likelihood):

$$-\sum_{i=1}^{N}[y_i \log \sigma(w_0 + \mathbf{w}^T\mathbf{x}_i) + (1 - y_i)\log(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{x}_i))] =$$
$$-\sum_{i=1}^{N}[y_i \log \sigma(w_0 + \mathbf{w}^T\mathbf{a} + b\mathbf{w}^T(\frac{\mathbf{x_i} - \mathbf{a}}{b})) + (1 - y_i)\log(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{a} + b\mathbf{w}^T(\frac{\mathbf{x_i} - \mathbf{a}}{b})))],$$

which means, similar to linear regression, we can optimize the loss over the transformed dataset by setting $w'_j = b \cdot w_j$ (for $j \neq 0$).
For hard-margin SVM we can rewrite the objective function as $\frac{1}{2}\|b\mathbf{w}\|^2$ and the constraints as:

$$y_i(w_0 + \mathbf{w}^T\mathbf{x}_i)) \geqslant 1 \Rightarrow y_i(w_0 + \mathbf{w}^T\mathbf{a} + b\mathbf{w}^T(\frac{\mathbf{x_i} - \mathbf{a}}{b})) \geqslant 1.$$

Hence, to solve the optimization problem for the transformed dataset we can set $w'_j = b \cdot w_j$ (for $j \neq 0$).
Regarding the linear separability: since the transformation is linear, it won't change the linear separability of the data.

(f) $H_1 \subseteq H_2 \Rightarrow$ VC-dim$(H_1) \leqslant$ VC-dim(H_2), where H_1 and H_2 are two finite hypothesis classes.

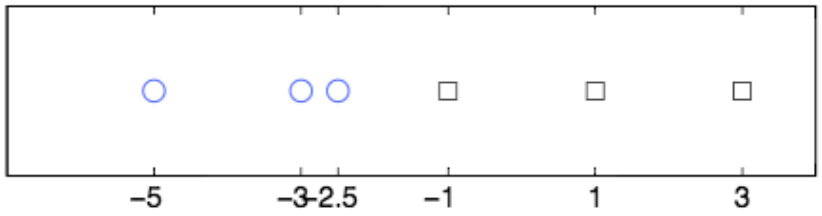
- ◯ True
- ◯ False

Explain your reasoning in 1-3 sentences. Just answering True or False gets zero points.

Solution: True. Let VC-dim$(H_1) = d$. Consider a set S of d points that can be shattered by H_1. Since every $h \in H_1$ required to shatter S is also in H_2 (because H_2 contains H_1), H_2 can also shatter S, therefore VC-dim$(H_2) \geqslant d =$ VC-dim(H_1).

(g) Consider the following one-dimensional data set:

Name: _____



1. Where is the decision boundary of a (hard-margin) linear SVM classifier trained on the full data set?

   Solution: $x = -1.75$

2. What is the leave-one-out cross-validation error (i.e. fraction of misclassified points) for a hard-margin linear SVM classifer on this data? "Leave-one-out" cross-validation means k-fold cross-validation with k equal to the number of data points (i.e., 6).

   Solution: $1/6$

3. What is the leave-one-out cross-validation error for a 1-NN classifier on this data?

   Solution: $1/6$

Name: _____

Let us Do Regression

2. (34 points) For the purpose of this question, we will consider data being generated as per a *mixture model*. Specifically, let $X \in \mathbb{R}$ denote the random variable representing the features and $Y \in \mathbb{R}$ denote the target or label that we would like to *predict* using the features.

   We assume that random variable X is generated as per mixture of m-Gaussian distributions with ith mixture component being Gaussian with mean $\mu_i \in \mathbb{R}$ and variance $\sigma_i^2 \geqslant 0$, and the probability of ith mixture being p_i with $1 \leqslant i \leqslant m$. To put it another way, to generate a sample of random variable X:

   - we sample random variable π which has multinomial distribution on $\{1, \ldots, m\}$ such that $\mathbb{P}(\pi = i) = p_i$, for $1 \leqslant i \leqslant m$
   - if the outcome of the multinomial is i, that is $\pi = i$, then we generate a sample from a Gaussian distribution with mean μ_i and variance σ_i^2.

   Given X, we generate Y as follows: if $\pi = i$, then $Y = w_i X + \epsilon$, where $w_i \in \mathbb{R}$ is a fixed parameter associated with mixture component i and ϵ is independent Gaussian with mean 0 and variance 1.

   (a) Let us make sure that we understand the setup by answering few simple questions.

   1. What is the probability density function of X? *Note*: the probability density function of a Gaussian distribution with mean μ and variance σ^2 is $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

   Solution:

   $$\mathbb{P}(X = x) = \sum_{i=1}^{m} \mathbb{P}(X = x | \pi = i)\mathbb{P}(\pi = i) = \sum_{i=1}^{m} \frac{p_i}{\sqrt{2\pi}\sigma_i} \exp(\frac{-1}{2\sigma_i^2}(x - \mu_i)^2)$$

   2. Suppose you observe that $X = x$, what is the likelihood that it came from mixture component i? That is, compute $\mathbb{P}(\pi = i | X = x)$.

   Solution:

   $$\mathbb{P}(\pi = i | X = x) = \frac{\mathbb{P}(\pi = i, X = x)}{\mathbb{P}(X = x)} = \frac{\frac{p_i}{\sigma_i} \exp(\frac{-1}{2\sigma_i^2}(X - \mu_i)^2)}{\sum_{j=1}^{m} \frac{p_j}{\sigma_j} \exp(\frac{-1}{2\sigma_j^2}(X - \mu_j)^2)}$$

   (b) Next, let us suppose that you know that data is generated as per the above described distribution. Let $f^* : \mathbb{R} \to \mathbb{R}$ be such that it minimizes $\mathbb{E}[(Y - f(X))^2]$ over all choices of function $f : \mathbb{R} \to \mathbb{R}$. Please provide an explicit form of f^* using your knowledge of the joint distribution of X, Y.

   Note: if you didn't work out $\mathbb{P}(\pi = i | X = x)$ in (a).2, you can use the notation $\mathbb{P}(\pi = i | X = x)$ in your solution for (b) instead of its explicit form.

Solution:

$$f^* = \mathbb{E}(Y|X) = \mathbb{E}_{\pi|X}[\mathbb{E}(Y|\pi, X)] = \sum_{i=1}^{m} \mathbb{P}(\pi = i|X)\mathbb{E}(Y|\pi = i, X) = \sum_{i=1}^{m} \frac{\mathbb{P}(\pi = i, X)}{\mathbb{P}(X)} w_i X$$

$$= \frac{\sum_{i=1}^{m} \frac{p_i}{\sigma_i} \exp(\frac{-1}{2\sigma_i^2}(X - \mu_i)^2) w_i X}{\sum_{j=1}^{m} \frac{p_j}{\sigma_j} \exp(\frac{-1}{2\sigma_j^2}(X - \mu_j)^2)}$$

(c) Now, suppose we do not have the knowledge of the joint distribution of X, Y. But we observe N data points, (x_n, y_n) for $1 \leqslant n \leqslant N$. Using these observations, we would like to identify f^*. As a start, we would like to understand whether simple linear regression would be a good idea or not. That is, we want to find a function $f(x) = a_N^* x + b_N^*$ so that (a_N^*, b_N^*) are solutions to

$$\text{minimize} \sum_{n=1}^{N} (y_n - a x_n - b)^2 \quad \text{over} \quad a, b \in \mathbb{R}.$$

Identify limiting quantities, $a_\infty^* = \lim_{N\to\infty} a_N^*$ as well as $b_\infty^* = \lim_{N\to\infty} b_N^*$. You may write your results using statistical quantities of X and Y (e.g. expectation, variance, etc).

Solution: This is a simple linear regression problem, the solutions for a_N^* and b_N^* are given by

$$a_N^* = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \; b_N^* = \bar{y} - a_N^* \bar{x}.$$

When $N \to \infty$,

$$\bar{x} \to \mathbb{E}(X), \; \bar{y} \to \mathbb{E}(Y), \; \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \to \text{Cov}(X, Y) \text{ and } \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \to \text{Var}(X).$$

So we have

$$a_\infty^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{ and } b_\infty^* = \mathbb{E}(Y) - a_\infty^* \mathbb{E}(X).$$

(d) Given that we understand a_∞^* and b_∞^*, let us examine the performance of linear regression (assuming $N = \infty$). To that end, let us consider a concrete scenario where $m = 2$, $p_1 = p_2 = 1/2$, $w_1 = 1$, $w_2 = -1$, $\mu_1 = \theta$, $\mu_2 = -\theta$ for some $\theta > 0$ and $\sigma_1 = \sigma_2 = 1$.
What are the values of a_∞^* and b_∞^* for this specific setting? Will the linear estimator trained with infinitely many observations yield good predictions for this setting (yes/no)?

Solution: We have

$$\mathbb{E}(XY) = \mathbb{E}_\pi \mathbb{E}(XY|\pi) = \frac{1}{2}\mathbb{E}(XY|\pi = 1) + \frac{1}{2}\mathbb{E}(XY|\pi = 2) = \frac{1}{2}\mathbb{E}(X^2|\pi = 1) + \frac{1}{2}\mathbb{E}(-X^2|\pi = 2) = 0.$$

Name: _____

Since $\mathbb{E}(X) = 0$, we have $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$. Hence

$$a_\infty^* = 0 \text{ and } b_\infty^* = \mathbb{E}(Y) = \theta.$$

This means that linear estimator is not a good choice even with infinitely many observations.

Name: _____

Click-through rate (CTR) classification with perceptron and SVM

3. (34 points) Consider the CTR prediction problem from lecture. Our goal is to predict whether or not a user will click a given ad using an SVM classifier. An ad is represented as a subset, X, of d finite advertising techniques (e.g. whether the ad contains a celebrity, is blinking, uses neon colors, etc.) We can naturally represent X as a binary vector $\phi(x) \in \{0,1\}^d$ where the j-th dimension of $\phi(x)$ is 1 if the ad contains technique j and 0 otherwise. The labels y are either 1 (ad clicked) or -1 (not clicked).

   (a) Before using an SVM, let's first consider a perceptron classifier. We introduce η to be our learning rate for updating our perceptron weights. Let $\phi(x_i)$ be a feature representation of the ad set X_i and $y_i \in \{-1, +1\}$, a binary training label. For simplicity, do **not** include the bias weight, w_0, in your answers.

      1. Assume that at time t, current weights w^t misclassify ad X_i. Write the perceptron update rule in the form $w^{t+1} = w^t + ???$, where the missing term should be written in terms of η, $\phi(x_i)$, and y_i.

         Solution: $w^{t+1} = w^t + \eta y_i \phi(x_i)$

      2. Notice the perceptron update rule in (1) resembles stochastic gradient descent (SGD). Write the loss function for SGD with learning rate η that gives an update that is identical to the perceptron update for any ad (when the perceptron misclassifies). Loss functions are non-negative, so write your answer in the form $\mathcal{L}(w) = \max(0, ???)$ for a particular $\phi(x_i)$ and y_i.

         Solution: $\mathcal{L}(w) = \max(0, -y_i w^T \phi(x_i))$ (note the negative sign). To solve for the (???) term, set the SGD update ($-\eta \frac{\partial \mathcal{L}(w)}{\partial w}$) equal to the perceptron update in (1) and solve for $\mathcal{L}(w)$. This is for the case where the perceptron makes a mistake. If the perceptron does not make a mistake, there is no update, and $\mathcal{L}(w) = 0$. Grading note: due to the way this question is asked (in that we ask students to derive using (1)), we should accept solutions that are just $\mathcal{L}(w) = -y_i w^T \phi(x_i)$.

      3. Suppose we use the perceptron loss, $\mathcal{L}(w)$, from (2) in the loss-formulation of SVM using SGD (no margin constraints) to learn w. Compare this method (SVM with SGD) to perceptron, assuming the same hyper-parameters (e.g. learning rate) are used for both. Select the minimum number of boxes that will make the two methods identical.
         - ☐ There is no regularization term in SVM.
         - ☐ The training data is linearly separable.
         - ☐ The training data is presented in the same order for both methods.
         - ☐ The training data are all support vectors.
         - ☐ None of these and no other conditions are needed. They are already identical.

Solution: The first and third boxes should be checked. (See page 53 of lecture 7 for slides on the loss-formulation of SVM). The previous question shows us that perceptron can also be formulated as a SGD minimization task with loss, $\mathcal{L}(w)$. In the loss-formulation of SVM without margin constraints there is no notion of slack. Using the same $\mathcal{L}(w)$ loss as perceptron, if there is **no regularization**, then the SGD minimization task is the same as the perceptron as long as same the same terms are summed up to comprise w^*. This occurs when **SGD is trained in the same order on both**, so that both methods will see "mistakes" on the same set of points. Because $\mathcal{L}(w) = 0$ when examples are correctly classified, SGD only updates for mistakes, just like perceptron. Because both methods are minimizing the same loss, using the same method (SGD) with the same hyper-parameters and the exact same updates, the two methods are identical. This is true regardless of the support vectors or linear separability of the data.

4. Let α_i be an integer that represents how many times ad X_i has been misclassified and used to update the weights during training. Upon convergence, what is the prediction rule of the trained perceptron for a new ad, X, represented by $\phi(x)$? Assume the weights are initialized to zero. Your answer should be of the form $\hat{y}$ =??? and include the kernel function $k(X_i, X) = \langle \phi(x_i), \phi(x) \rangle$. Hint: Look at the answer to (1).

Solution: $\hat{y} = \text{sign}(\langle w^*, \phi(x) \rangle) = \text{sign}\left(\eta \sum_i \alpha_i y_i k(X_i, X)\right)$ Note that η does not affect the sign and is not necessary in the solution.

5. To formulate the *kernel* perceptron, we must replace (1) with an update rule that does not require an explicit representation of $\phi(x_i)$. For a misclassified ad X_i, write a new update of the form α_i = ??? such that $\hat{y}$ in (4) remains unchanged.

Solution: $\alpha_i = \alpha_i + 1$

(b) For both kernel-perceptron and SVM, our prediction $\hat{y}$ is a function of our kernel $K(X, Z)$ for two ads X and Z. For each of the following functions, state whether or not it is a valid kernel function. Be sure to show how you arrive at your answer.

1. Union (where |Q| denotes the number of elements in set Q):

$$k(X, Z) = |X \cup Z|$$

Solution: No. One sample counterexample: Consider the sets $X = \{\text{“celebrity”}\}$ and $Z = \{\text{“blinking”}\}$. Their Gram (kernel) matrix will be given by $K = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$. The eigenvalues of K are -1 and 3, so K is not positive semidefinite.

There are other valid explanations, such as showing that the Cauchy-Schwarz inequality does not hold.

2. Product:

$$k(X, Z) = |X| \cdot |Z|$$

Solution: Yes. Let $\phi(X) = |X|$.

3. Product + Intersection Cubed:

$$k(X, Z) = (\ |X| \cdot |Z| + |X \cap Z|\ )^3$$

Solution: Yes. The sum of kernels is a kernel and the product of kernels is a kernel. From the above question, $|X| \cdot |Z|$ is a kernel. $|X \cap Z|$ is a kernel, with $\phi(X) \in \{0, 1\}^d$ being the binary vector representation of X.

(c) Recall that in the soft-margin SVM (aka C-SVM), the objective function is to minimize $\frac{1}{2}||w||^2 + C \sum_i \xi_i$, where ξ_i is the slack for each training example. For any non-negative C, if we increase C, which of the following are possible effects on the value $||w||^2$:

( ) increases, ( ) stays the same, ( ) decreases

Solution: $||w||^2$ either increases or stays the same. In general: larger C $\rightarrow$ Less slack $\rightarrow$ Smaller margin $\rightarrow$ Larger w. However, in some regions, there will be no change in w. For example, if the data is linearly separable and C is very large, all points will be correctly classified, so increasing C further will not make any difference.

(d) Suppose we have the following training points:

Ad Number	Set of Advertising Techniques	Click?
1	{promotion, celebrity, blinking}	+1
2	{promotion, neon_color}	-1
3	{celebrity, neon_color, blinking}	+1
4	{blinking}	-1

Assume we train an SVM using the kernel $k(X, Z) = |X \cap Z|$ (e.g. $k(X, Z) = 2$ if sets X and Z have two common elements) and get values for $\alpha_1, ..., \alpha_4$. Suppose we are given a new test ad using the set of techniques {promotion, blinking}. Write the prediction (+1 or -1) for whether the new ad will be clicked, in terms of the α_i parameters.

Name: _____

Solution:

$$\hat{y} = \text{sign}[\sum_{i=1}^{4} \alpha_i y_i k(x_i, x)]$$

$$= \text{sign}[\alpha_1(1)(2) + \alpha_2(-1)(1) + \alpha_3(1)(1) + \alpha_4(-1)(1)]$$

$$= \text{sign}[2\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4]$$