

QUIZ2

- Recall that the probabilistic SVM is based on solving the following optimization problem:

$$\min_{A,B} F(A,B) = \frac{1}{N} \sum_{n=1}^N \ln(1 + \exp(-y_n (A \cdot (\mathbf{w}_{svm}^T \phi(\mathbf{x}_n) + b_{svm}) + B)))$$

When using the gradient descent for minimizing $F(A,B)$, we need to compute the gradient first. $z_n = \mathbf{w}_{svm}^T \phi(\mathbf{x}_n) + b_{svm}$, and $p_n = \theta(-y_n(Az_n + B))$, where $\theta(s) = \frac{\exp(s)}{1+\exp(s)}$ is the usual logistic function. What is the gradient $\nabla F(A,B)$?

- $\frac{1}{N} \sum_{n=1}^N [-y_n p_n z_n, -y_n p_n]^T$
 - $\frac{1}{N} \sum_{n=1}^N [-y_n p_n z_n, +y_n p_n]^T$
 - $\frac{1}{N} \sum_{n=1}^N [+y_n p_n z_n, -y_n p_n]^T$
 - $\frac{1}{N} \sum_{n=1}^N [+y_n p_n z_n, +y_n p_n]^T$
 - none of the other choices
- When using the Newton method for minimizing $F(A,B)$ (see Homework 3 of Machine Learning Foundations), we need to compute $-(H(F))^{-1} \nabla F$ in each iteration, where $H(F)$ is the Hessian matrix of F at (A,B) . Following the notations of Question 1, what is $H(F)$?

- $\frac{1}{N} \sum_{n=1}^N \begin{bmatrix} z_n^2 y_n (1 - p_n) & z_n y_n (1 - p_n) \\ z_n y_n (1 - p_n) & y_n (1 - p_n) \end{bmatrix}$
- $\frac{1}{N} \sum_{n=1}^N \begin{bmatrix} z_n^2 p_n (1 - y_n) & z_n p_n (1 - y_n) \\ z_n p_n (1 - y_n) & p_n (1 - y_n) \end{bmatrix}$
- none of the other choices
- $\frac{1}{N} \sum_{n=1}^N \begin{bmatrix} z_n^2 p_n (1 - p_n) & z_n p_n (1 - p_n) \\ z_n p_n (1 - p_n) & p_n (1 - p_n) \end{bmatrix}$
- $\frac{1}{N} \sum_{n=1}^N \begin{bmatrix} z_n^2 y_n (1 - y_n) & z_n y_n (1 - y_n) \\ z_n y_n (1 - y_n) & y_n (1 - y_n) \end{bmatrix}$

- Recall that N is the size of the data set and d is the dimensionality of the input space. What is the size of matrix that gets inverted in kernel ridge regression?
 - $d \times d$
 - $N \times N$
 - $Nd \times Nd$
 - $N^2 \times N^2$
 - none of the other choices

- The usual support vector regression model solves the following optimization problem.

$$(P_1) \quad \min_{b, \mathbf{w}, \xi_n^\vee, \xi_n^\wedge} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\xi_n^\vee + \xi_n^\wedge)$$

$$\text{s.t. } -\xi_n^\vee \leq y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b \leq \epsilon + \xi_n^\wedge \xi_n^\vee \geq 0, \xi_n^\wedge \geq 0.$$

Usual support vector regression penalizes the violations ξ_n^\vee and ξ_n^\wedge linearly. Another popular formulation, called ϵ -insensitive support vector regression in (P2), penalizes the violations quadratically, just like the ϵ -insensitive SVM introduced in Homework 1 of Machine Learning Techniques.

$$(P_2) \quad \min_{b, \mathbf{w}, \xi_n^\vee, \xi_n^\wedge} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \left((\xi_n^\vee)^2 + (\xi_n^\wedge)^2 \right)$$

$$\text{s.t. } -\xi_n^\vee \leq y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b \leq \epsilon + \xi_n^\wedge.$$

Which of the following is an equivalent ‘unconstrained’ form of (P2)?

- A. none of the other choices
- B. $\min_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (|y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon)^2$
- C. $\min_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\max(0, |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon))^2$
- D. $\min_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\max(\epsilon, |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b|))^2$
- E. $\min_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b)^2$

5. By a slight modification of the representer theorem presented in the class, the optimal \mathbf{w}_* for (P2) must satisfy $\mathbf{w}_* = \sum_{n=1}^N \beta_n \mathbf{z}_n$. We can substitute the form of the optimal \mathbf{w}_* into the answer in Question 4 to derive an optimization problem that contains β (and b) only, which would look like

$$\min_{b, \beta} F(b, \beta) = \frac{1}{2} \sum_{m=1}^N \sum_{n=1}^N \beta_n \beta_m K(\mathbf{x}_n, \mathbf{x}_m) + \text{something},$$

where $K(\mathbf{x}_n, \mathbf{x}_m) = (\phi(\mathbf{x}_n))^T (\phi(\mathbf{x}_m))$ is the kernel function. One thing that you should see is that $F(b, \beta)$ is differentiable to β_n (and b) and hence you can use gradient descent to solve for the optimal β . For any β , let $s_n = \sum_{m=1}^N \beta_m K(\mathbf{x}_n, \mathbf{x}_m) + b$. What is $\frac{\partial F(b, \beta)}{\partial \beta_m}$?

- A. $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}_m) - 2C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] (|y_n - s_n| - \epsilon) \text{sign}(y_n - s_n) K(\mathbf{x}_n, \mathbf{x}_m)$
- B. $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}_m) + 2C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] (|y_n - s_n| - \epsilon) \text{sign}(y_n - s_n) K(\mathbf{x}_n, \mathbf{x}_m)$
- C. $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}_m) - 2C \sum_{n=1}^N [|y_n - s_n| \leq \epsilon] (|y_n - s_n| - \epsilon) \text{sign}(y_n - s_n) K(\mathbf{x}_n, \mathbf{x}_m)$
- D. $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}_m) + 2C \sum_{n=1}^N [|y_n - s_n| \leq \epsilon] (|y_n - s_n| - \epsilon) \text{sign}(y_n - s_n) K(\mathbf{x}_n, \mathbf{x}_m)$
- E. none of the other choices

6. Consider $T+1$ hypotheses g_0, g_1, \dots, g_T . Let $g_0(\mathbf{x}) = 0$ for all x . Assume that your boss holds a test set $\{(\tilde{\mathbf{x}}_m, \tilde{y}_m)\}_{m=1}^M$, where you know $\tilde{\mathbf{x}}_m$ but \tilde{y}_m is hidden. Nevertheless, you are allowed to know the squared test error $E_{\text{test}}(g_t) = \frac{1}{M} \sum_{m=1}^M (g_t(\tilde{\mathbf{x}}_m) - \tilde{y}_m)^2 = e_t$ for $t = 0, 1, 2, \dots, T$. Also, assume that $\frac{1}{M} \sum_{m=1}^M (g_t(\tilde{\mathbf{x}}_m))^2 = s_t$. Which of the following allows you to calculate $\sum_{m=1}^M g_t(\tilde{\mathbf{x}}_m) \tilde{y}_m$? Note that the calculation is the key to the test set blending technique that the NTU team has used in KDDCup2011.

- A. $\frac{M}{2} (-e_0 - s_t + e_t)$
- B. $\frac{M}{2} (+e_0 - s_t + e_t)$
- C. $\frac{M}{2} (+e_0 + s_t - e_t)$
- D. none of the other choices
- E. $\frac{M}{2} (-e_0 + s_t - e_t)$

7. Consider the case where the target function $f : [0, 1] \rightarrow \mathbb{R}$ is given by $f(x) = x^2$ and the input probability distribution is uniform on $[0, 1]$. Assume that the training set has only two examples generated independently from the input probability distribution and noiselessly by f , and the learning model is usual linear regression that minimizes the mean squared error within all hypotheses of the form $h(x) = w_1 x + w_0$. What is $\bar{g}(x)$, the expected value of the hypothesis that the learning algorithm produces (see Page 10 of Lecture 207)?

- A. $\bar{g}(x) = 2x - \frac{1}{2}$
- B. $\bar{g}(x) = 2x + \frac{1}{2}$
- C. $\bar{g}(x) = x - \frac{1}{4}$
- D. $\bar{g}(x) = x + \frac{1}{4}$
- E. none of the other choices

8. Assume that linear regression (for classification) is used within AdaBoost. That is, we need to solve the weighted- E_{in} optimization problem

$$\min_{\mathbf{w}} E_{in}^{\mathbf{u}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N u_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2.$$

- A. none of the other choices
- B. $(\sqrt{u_n} \mathbf{x}_n, \sqrt{u_n} y_n)$
- C. $(u_n^{-2} \mathbf{x}_n, u_n^{-2} y_n)$
- D. $(u_n^2 \mathbf{x}_n, u_n^2 y_n)$
- E. $(u_n \mathbf{x}_n, u_n y_n)$

9. Consider applying the AdaBoost algorithm on a binary classification data set where 99% of the examples are positive. Because there are so many positive examples, the base algorithm within AdaBoost returns a constant classifier $g_1(\mathbf{x}) = +1$ in the first iteration. Let $u_+^{(2)}$ be the individual example weight of each positive example in the second iteration, and $u_-^{(2)}$ be the individual example weight of each negative example in the second iteration. What is $u_+^{(2)}/u_-^{(2)}$?

- A. none of the other choices
- B. 1/100
- C. 1/99
- D. 100
- E. 99

10. When talking about non-uniform voting in aggregation, we mentioned that can be viewed as a weight vector learned from any linear algorithm coupled with the following transform:

$$\phi(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_T(\mathbf{x})).$$

When studying kernel models, we mentioned that the kernel is simply a computational short-cut for the inner product $(\phi(\mathbf{x}))^T \phi(\mathbf{x}')$. In this problem, we mix the two topics together using the decision stumps as our $g_t(\mathbf{x})$.

Assume that the input vectors contain only integers between (including) L and R .

$$g_{s,i,\theta}(\mathbf{x}) = s \cdot \text{sign}(x_i - \theta),$$

where $i \in \{1, 2, \dots, d\}$, d is the finite dimensionality of the input space,

$$s \in \{-1, +1\}, \theta \in \mathbb{R}, \text{ and } \text{sign}(0) = +1$$

Two decision stumps g and \hat{g} are defined as the same if $g(\mathbf{x}) = \hat{g}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Two decision stumps are different if they are not the same. Which of the followings are true?

- A. The number of different decision stumps equals the size of \mathcal{X}
- B. \mathcal{X} is of infinite size
- C. There are 22 different decision stumps for the case of $d = 2$, $L = 1$, and $R = 6$

- D. $g_{+1,1,L-1}$ is the same as $g_{-1,3,R+1}$
- E. $g_{s,i,\theta}$ is the same as $g_{s,i,\text{ceiling}(\theta)}$, where $\text{ceiling}(\theta)$ is the smallest integer that is greater than or equal to θ
11. Continuing from the previous question, let $\mathcal{G} = \{ \text{all different decision stumps for } \mathcal{X} \}$ and enumerate each hypothesis $g \in \mathcal{G}$ by some index t . Define

$$\phi_{ds}(\mathbf{x}) = \left(g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_t(\mathbf{x}), \dots, g_{|\mathcal{G}|}(\mathbf{x}) \right).$$

Derive a simple equation that evaluates $K_{ds}(\mathbf{x}, \mathbf{x}') = (\phi_{ds}(\mathbf{x}))^T (\phi_{ds}(\mathbf{x}'))$ efficiently. Which of the following equation is correct? Here $\|\mathbf{v}\|_1$ denotes the one-norm of \mathbf{v} .

- A. $K_{ds}(\mathbf{x}, \mathbf{x}') = 2d(R - L) - 4\|\mathbf{x} - \mathbf{x}'\|_1 - 2$
- B. none of the other choices
- C. $K_{ds}(\mathbf{x}, \mathbf{x}') = 2d(R - L) - 4\|\mathbf{x} - \mathbf{x}'\|_1 + 2$
- D. $K_{ds}(\mathbf{x}, \mathbf{x}') = d(R - L) - 2\|\mathbf{x} - \mathbf{x}'\|_1 - 2$
- E. $K_{ds}(\mathbf{x}, \mathbf{x}') = d(R - L) - 2\|\mathbf{x} - \mathbf{x}'\|_1 + 2$
12. For Questions 12-18, implement the AdaBoost-Stump algorithm as introduced in Lecture 208. Run the algorithm on the following set for training: *hw2_adaboost_train.dat* and the following set for testing: *adaboost_test.dat*. Use a total of $T = 300$ iterations (please do not stop earlier than 300), and calculate E_{in} and E_{out} with the 0/1 error. For the decision stump algorithm, please implement the following steps. Any ties can be arbitrarily broken.
1. For any feature i , sort all the $x_{n,i}$ values to $x_{[n],i}$ such that $x_{[n],i} \leq x_{[n+1],i}$.
 2. Consider thresholds within $-\infty$ and all the midpoints $\frac{x_{[n],i} + x_{[n+1],i}}{2}$. Test those thresholds with $s \in \{-1, +1\}$ to determine the best (s, θ) combination that minimizes E_{in}^u using feature i .
 3. Pick the best (s, i, θ) combination by enumerating over all possible i .
- For those interested, Step 2 can be carried out in $O(N)$ time only!!

Which of the following is true about $E_{in}(g_1)$?

- A. $0.2 \leq E_{in}(g_1) < 0.3$
- B. $E_{in}(g_1) > 0.3$
- C. $E_{in}(g_1) = 0$
- D. 0
- E. $0.1 \leq E_{in}(g_1) < 0.2$
13. Which of the following is true about $E_{in}(G)$?
- A. $0.1 \leq E_{in}(G) < 0.2$
- B. $0.2 \leq E_{in}(G) < 0.3$
- C. $E_{in}(G) > 0.3$
- D. $0 < E_{in}(G) < 0.1$
- E. $E_{in}(G) = 0$
14. Let $U_t = \sum_{n=1}^N u_n^{(t)}$. Which of the following is true about U_2 ? (note that $U_1 = 1$)
- A. $U_2 = 0$
- B. $0 < U_2 < 0.1$
- C. $0.1 \leq U_2 < 0.2$

- D. $0.2 \leq U_2 < 0.3$
 E. $U_2 > 0.3$
15. Which of the following is true about U_T ?
- A. $U_T = 0$
 B. $0 < U_T < 0.1$
 C. $0.1 \leq U_T < 0.2$
 D. $0.2 \leq U_T < 0.3$
 E. $U_T > 0.3$
16. Which of the following is true about the minimum value of ϵ_t within $t = 1, 2, \dots, 300$?
- A. $0 < \text{value} < 0.1$
 B. $\text{value} > 0.3$
 C. $\text{value} = 0$
 D. $\text{value} = 0$
 E. $0.2 \leq \text{value} < 0.3$
17. Calculate E_{out} with the test set. Which of the following is true about $E_{out}(g_1)$?
- A. $0.2 \leq E_{out}(g_1) < 0.3$
 B. $E_{out}(g_1) > 0.3$
 C. $0 < E_{out}(g_1) < 0.1$
 D. $0.1 \leq E_{out}(g_1) < 0.2$
 E. $E_{out}(g_1) = 0$
18. Which of the following is true about $E_{out}(G)$?
- A. $0.1 \leq E_{out}(G) < 0.2$
 B. $0 < E_{out}(G) < 0.1$
 C. $E_{out}(G) = 0$
 D. $E_{out}(G) > 0.3$
 E. $0.2 \leq E_{out}(G) < 0.3$
19. Write a program to implement the kernel ridge regression algorithm from Lecture 206, and use it for classification (i.e. implement LSSVM). Consider the following data set [hw2_lssvm_all.dat](#). Use the first 400 examples for training and the remaining for testing. Calculate E_{in} and E_{out} with the 0/1 error. Consider the Gaussian-RBF kernel $\exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$. Try all combinations of parameters $\gamma \in \{32, 2, 0.125\}$ and $\lambda \in \{0.001, 1, 1000\}$. Among all parameter combinations, which of the following is the range that the minimum $E_{in}(g)$ resides in?
- A. $[0.8, 1.0)$
 B. $[0, 0.2)$
 C. $[0.4, 0.6)$
 D. $[0.2, 0.4)$
 E. $[0.6, 0.8)$

20. Following Question 19, among all parameter combinations, which of the following is the range that the minimum $E_{out}(g)$ resides in?
- A. $[0.2, 0.4)$
 - B. $[0.8, 1.0)$
 - C. $[0.4, 0.6)$
 - D. $[0.6, 0.8)$
 - E. $[0, 0.2)$