Homework #7

TA in charge: Yu-Xun Ruan and Chen-Wei Hung

RELEASE DATE: 12/13/2010

DUE DATE: 12/27/2010, 4:00 pm IN CLASS

TA SESSION: 12/23/2010, 6:00 pm IN R110

Unless granted by the instructor in advance, you must turn in a hard copy of your solutions (without the source code) for all problems. For problems marked with (*), please follow the guidelines on the course website and upload your source code to designated places.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

7.1 Kernels and Transforms

(1) (5%) Let all the input vectors be real values ($\mathbf{x} \in \mathcal{X} \in \mathbb{R}$) with $|\mathbf{x}| < 1$. Consider a kernel function $K(\mathbf{x}, \mathbf{x}') = \frac{1}{1-\mathbf{x} \bullet \mathbf{x}'}$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \bullet \boldsymbol{\phi}(\mathbf{x}').$$

(2) (5%) For two valid kernels $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x}) \bullet \boldsymbol{\phi}_1(\mathbf{x}')$ and $K_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_2(\mathbf{x}) \bullet \boldsymbol{\phi}_2(\mathbf{x}')$, consider a kernel function $K(\mathbf{x}, \mathbf{x}') = \gamma_1 K_1(\mathbf{x}, \mathbf{x}') + \gamma_2 K_2(\mathbf{x}, \mathbf{x}')$ with $\gamma_1 > 0$ and $\gamma_2 > 0$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \bullet \boldsymbol{\phi}(\mathbf{x}').$$

(The result shows that a conic combination of valid kernels is still a valid kernel.)

(3) (5%) For two valid kernels $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x}) \bullet \boldsymbol{\phi}_1(\mathbf{x}')$ and $K_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_2(\mathbf{x}) \bullet \boldsymbol{\phi}_2(\mathbf{x}')$, consider a kernel function $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \bullet \boldsymbol{\phi}(\mathbf{x}').$$

(The result shows that a multiplication of valid kernels is still a valid kernel.)

(4) (5%) For a given data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}$ in which all $\mathbf{x}_n$ are different. Consider the feature transform in Problem 5.2(1):

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{cases} (\underbrace{0, \cdots, 0}_{n-1}, 1, 0, \cdots) & \text{if } \mathbf{x} = \mathbf{x}_n \\ (0, \cdots, 0, 0, 0, \cdots) & \text{otherwise .} \end{cases}$$

What is the kernel function that corresponds to this transform? When using the kernel function, what does the matrix Q in the dual problem of SVM look like?

(5) (5%) Prove that the function $K(\mathbf{x}, \mathbf{x}') = -(\mathbf{x} \bullet \mathbf{x}')^2 + 2\mathbf{x} \bullet \mathbf{x}'$ is NOT a valid kernel function.

7.2 Kernel from Decision Stumps

When talking about non-uniform voting in aggregation, we mentioned that $\boldsymbol{\alpha}$ can be viewed as a weight vector learned from any linear algorithm coupled with the following transform:

$$\boldsymbol{\phi}(\mathbf{x}) = \Big(h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_T(\mathbf{x})\Big).$$

When studying kernel methods, we mentioned that the kernel is simply a computational short-cut for the inner product $\boldsymbol{\phi}(\mathbf{x}) \bullet \boldsymbol{\phi}(\mathbf{x}')$. In this problem, we mix the two topics together using the decision stumps as our $h(\mathbf{x})$.

(1) (5%) Assume that the input vectors contain only bounded integers. That is, $\mathcal{X} \subseteq (\mathbb{Z} \cap [-B, B])^d$. Consider decision stumps

$$h_{s,i,\theta}(\mathbf{x}) = \text{sign}\Big(s \cdot \mathbf{x}[i] - \theta\Big).$$

Two decision stumps $h^{(1)}$ and $h^{(2)}$ are defined as the *same* if $h^{(1)}(\mathbf{x}) = h^{(2)}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Two decision stumps are different if they are not the same. Argue that there are only finitely-many different decision stumps for $\mathcal{X}$ and list all of them for the case of $d = 3$ and $B = 2$.

(2) (5%) Let $\mathcal{H} = \{$ all different decision stumps for $\mathcal{X}$ $\}$. Since $\mathcal{H}$ is finite, we can enumerate each hypothesis $h \in \mathcal{H}$ by some index t. Define

$$\boldsymbol{\phi}_{ds}(\mathbf{x}) = \Big(h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_t(\mathbf{x}), \cdots, h_{|\mathcal{H}|}(\mathbf{x})\Big).$$

Derive a simple equation that evaluates $K_{ds}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_{ds}(\mathbf{x}) \bullet \boldsymbol{\phi}_{ds}(\mathbf{x}')$ efficiently.

(3) (Bonus 5%) Argue that for any kernel function $K(\mathbf{x}, \mathbf{x}')$, using $K(\mathbf{x}, \mathbf{x}') + c$ for any constant c is equivalent to using $K(\mathbf{x}, \mathbf{x}')$ in the dual of (hard- or soft-margin) SVM. In other words, prove that the resulting classifier is exactly the same. Use your argument to obtain a kernel of the form $K_{ds}(\mathbf{x}, \mathbf{x}') + c$ that is even easier to evaluate.

The result can be easily extended to the case when $\mathcal{X}$ is an arbitrary box in $\mathbb{R}^d$ as well.

7.3 Power of Adaptive Boosting

The adaptive boosting (AdaBoost) algorithm, as shown in the class slides, is as follows:

- For input $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, set $u_n = \frac{1}{N}$ for all n.
- For $t = 1, 2, \cdots, T$,
  - Learn a simple hypothesis h_t such that h_t solves

$$h_t = \underset{h \in \mathcal{H}}{\text{argmin}} \sum_{n=1}^{N} u_n \cdot [\![y_n \neq h(\mathbf{x}_n)]\!].$$

    with the help of some base learner A_b that learns from $h \in \mathcal{H}$.
  - Compute the weighted error $\epsilon_t = \dfrac{\sum_{n=1}^N u_n \cdot [\![y_n \neq h_t(\mathbf{x}_n)]\!]}{\sum_{n=1}^N u_n}$ and the confidence

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}$$

  - Change the example weights: $u_n = u_n \cdot \exp\Big(-\alpha_t y_n h_t(\mathbf{x}_n)\Big)$.
- Output: combined function $H(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})\right)$

In this problem, we will prove that AdaBoost can reach $E_{\text{in}}(H) = 0$ if T is large enough and every hypothesis h_t satisfies $\epsilon_t \leq \epsilon < \frac{1}{2}$.

(1) (5%) Let $U^{(t-1)} = \sum_{n=1}^{N} u_n$ at the beginning of the t-th iteration. According to the AdaBoost algorithm above, for $t \geq 1$, prove that

$$U^{(t)} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{t} \alpha_\tau h_\tau(\mathbf{x}_n)\right).$$

(2) (5%) By the result in (2), prove that $E_{\text{in}}(H) \leq U^{(T)}$.

(3) (5%) According to the AdaBoost algorithm above, for $t \geq 1$, prove that $U^{(t)} = U^{(t-1)} \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)}$.

(4) (5%) Using $0 \leq \epsilon_t \leq \epsilon < \frac{1}{2}$, for $t \geq 1$, prove that $\sqrt{\epsilon_t(1-\epsilon_t)} \leq \sqrt{\epsilon(1-\epsilon)}$.

(5) (5%) Using $\epsilon < \frac{1}{2}$, prove that $\sqrt{\epsilon(1-\epsilon)} \leq \frac{1}{2}\exp\left(-2(\frac{1}{2}-\epsilon)^2\right)$.

(6) (5%) Using the results above, prove that $U^{(T)} \leq \exp\left(-2T(\frac{1}{2}-\epsilon)^2\right)$.

(7) (5%) Using the results above, argue that after $T = O(\log N)$ iterations, $E_{\text{in}}(H) = 0$.

7.4 Optimization View of Adaptive Boosting

Assume that $\mathcal{H} = \{h_\ell\}_{\ell=1}^{L}$ is a finite set of hypotheses. Consider an error function for an ensemble $H(\mathbf{x})$ of the form $\text{sign}\left(\sum_{\ell=1}^{L} \beta_\ell h_\ell(\mathbf{x})\right)$:

$$E_{\exp}(H) = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\ell=1}^{L} \beta_\ell h_\ell(\mathbf{x}_n)\right).$$

Using your results in Problem 7.4(3), we can easily show that $E_{\exp}(H)$ is an upper bound of $E_{\text{in}}(H)$. We now prove that AdaBoost is equivalent to a particular way of minimizing $E_{\exp}(H)$.

(1) (5%) Assume that we hope to update from $\boldsymbol{\beta}^{old}$ to $\boldsymbol{\beta}^{new}$ by changing only component i of $\boldsymbol{\beta}^{old}$. That is, for a given vector $\boldsymbol{\beta}^{old} = (\beta_1^{old}, \beta_2^{old}, \cdots, \beta_{L-1}^{old}, \beta_L^{old})$, we want to set

$$\beta_i^{new} = \beta_i^{old} + \Delta_i.$$

such that

$$\Delta_i = \underset{\Delta}{\text{argmin}} \ \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \left(\sum_{\ell=1}^{L} \beta_\ell^{old} h_\ell(\mathbf{x}_n)\right) - y_n\big(\Delta \cdot h_i(\mathbf{x}_n)\big)\right).$$

Let $u_n = \frac{1}{N}\exp\left(-y_n \sum_{\ell=1}^{L} \beta_\ell^{old} h_\ell(\mathbf{x}_n)\right)$ and $\epsilon_i = \frac{\sum_{n=1}^{N} u_n \cdot [\![y_n \neq h_i(\mathbf{x}_n)]\!]}{\sum_{n=1}^{N} u_n}$. What is the optimal Δ_i in terms of ϵ_i?

(The result shows that the α_t in AdaBoost is the steepest descent choice for $E_{\exp}$ after getting h_t.)

(2) (5%) Suppose now that we want to pick the best i that greedily makes $E_{\exp}(H)$ the smallest. That is, for a given vector $(\beta_1, \beta_2, \cdots, \beta_{L-1}, \beta_L)$, we want to solve

$$\min_{\Delta, i} \ \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \left(\sum_{\ell=1}^{L} \beta_\ell^{old} h_\ell(\mathbf{x}_n)\right) - y_n\big(\Delta \cdot h_i(\mathbf{x}_n)\big)\right).$$

If all $h \in \mathcal{H}$ satisfies

$$\frac{\sum_{n=1}^{N} u_n \cdot [\![y_n \neq h(\mathbf{x}_n)]\!]}{\sum_{n=1}^{N} u_n} \leq \frac{1}{2},$$

show that the optimal

$$h_i = \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{n=1}^{N} u_n \cdot [\![y_n \neq h(\mathbf{x}_n)]\!] .$$

(The result shows that the h_t in AdaBoost is the optimal coordinate choice for E_{exp}.)

All the results can be extended to the case when $\mathcal{H}$ is of an infinite size as well. This problem shows that AdaBoost is optimizing a particular error function E_{exp} slowly (by coordinate).

7.5 Experiments with Adaptive Boosting (*)

(1) (10%) Implement the AdaBoost algorithm with decision stumps (i.e., use A_{ds} as A_b). Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml10fall/data/hw7_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml10fall/data/hw7_test.dat`

Use a total of $T = 300$ iterations. Let $H_t(\mathbf{x}) = \text{sign}\left(\sum_{\tau=1}^{t} \alpha_\tau h_\tau(\mathbf{x})\right)$. Plot $E_{\text{in}}(H_t)$, $E_{\text{out}}(H_t)$ and $U^{(t)}$ (see the definition above) as functions of t on the same figure. Briefly state your findings.

7.6 Experiments with Unpruned Decision Tree (*)

(1) (10%) Implement the following simple decision tree algorithm (without pruning):

`function DecisionTree(`$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$`) returns a tree`

- When every example in $\mathcal{D}$ shares the same y_n, return a constant hypothesis $h(\mathbf{x}) = y_n$.
- Learn a decision-stump branch function $b(\mathbf{x})$ that separates $\mathcal{D}$ to 2 parts:

$$b(\mathbf{x}) = \operatorname*{argmax}_{h_{s,i,\theta}} \left(\text{impurity}(\mathcal{D}) - \sum_{c=1}^{2} \frac{|\mathcal{D}^{(c)}|}{|\mathcal{D}|} \text{impurity}(\mathcal{D}^{(c)}) \right)$$

  where $\mathcal{D}^{(c)} = \{(\mathbf{x}_n, y_n) \colon b(\mathbf{x}_n) = c\}$ and the impurity is measured by Gini index.
- Obtain $H_c^{(L-1)} =$ `DecisionTree(`$\mathcal{D}^{(c)}$`)`
- Return

$$H^{(L)}(\mathbf{x}) = \sum_{c=1}^{2} [\![b(\mathbf{x}) = c]\!] \cdot H_c^{(L-1)}(\mathbf{x})$$

Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml10fall/data/hw7_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml10fall/data/hw7_test.dat`

Plot the training examples $(\mathbf{x}_n, y_n)$ and the decision boundary (where the prediction changes) in the same figure. In addition, show your binary tree $H^{(L)}$ (by graph or by writing down the if-then-else). Report $E_{\text{in}}(H^{(L)})$ and $E_{\text{out}}(H^{(L)})$. Briefly state your findings.