Theoretical Machine Learning - COS 511

Homework Assignment 2

Due Date: 04 Apr 2017, till 22:00

(1) Solve 4 out of the following 6 problems.

(2) Consulting other students from this course is allowed. In this case - clearly state whom you consulted with for each problem separately.

(3) Searching the internet or literature for solutions, other than the course lecture notes, is NOT allowed.

(4) All problems are weighted equally at 10 points each. Indicate on your problem set which four problems you choose to solve. Feel free to write down solutions for the other two as well, but your homework grade will only depend upon the four you mark to be graded.

Ex. 1:

Prove or disprove by counter example: For every learning problem $(\mathcal{Z}, \mathcal{C}, \ell)$, an arbitrary ERM algorithm learns the class if and only if the class has the uniform convergence property. In short, consider an algorithm that receives a sample $S = \{\mathbf{z}^{(i)}\}_{i=1}^m$ and returns some $h \in \mathcal{C}$ such that

$$\sum_{i=1}^{m} \ell(h_S^A; \mathbf{z})) = \inf_{h^* \in \mathcal{C}} \sum_{i=1}^{m} \ell(h^*; \mathbf{z}))$$

Show (or disprove): For sufficiently large $m > m(\epsilon, \delta)$ we have w.p $1 - \delta$:

$$\mathbb{E}\left[\ell(h_S^A, \mathbf{z})\right] \leq \inf_{h \in \mathcal{C}} \mathbb{E}\left[\ell(h, \mathbf{z})\right] + \epsilon$$

if and only if $(\mathcal{Z}, \mathcal{C}, \ell)$ has the uniform convergence property.

Ex. 2:

In this exercise we are going to show an "inefficient" boosting result. Namely, we will prove existence of Boosting but without an algorithm. Let $(\mathcal{X}, \mathcal{H}, \ell_{0,1})$ be a binary classification problem (realizable). For simplicity assume $|\mathcal{X}| < \infty$. Assume that: for every distribution D there exists $h_D \in \mathcal{H}$ such that

$$\mathrm{err}(h_D) \leq \frac{1}{2} - \gamma.$$

We consider the class of classifiers

$$\bar{\mathcal{H}} = \{\bar{h} : \bar{h} = \mathrm{sign}(\sum \lambda_i h_i),\ \lambda_i \geq 0,\ \sum \lambda_i = 1,\ h_i \in \mathcal{H}\}$$

Let M be a matrix whose column corresponds to labeled elements in $\mathcal{X}$ and whose rows correspond to elements in $\mathcal{H}$, so that:

$$M(i,j) = \begin{cases} 0 & h_j(\mathbf{x}^{(i)}) = y_i \\ 1 & \text{else} \end{cases}$$

Show that under the assumptions above (of weak learnability) there exists a distribution $\mathbf{u} \in \Delta_{|\mathcal{H}|}$ over the hypothesis space such that for any distribution $\mathbf{v} \in \Delta_{|\mathcal{X}|}$ over the sample points:

$$\mathbf{v}^\top M \mathbf{u} < \frac{1}{2} - \gamma$$

Conclude that there exists $\bar{h} \in \bar{\mathcal{H}}$ such that $\mathrm{err}(\bar{h}) = 0$. You will need the following minmax statement:

The minmax theorem asserts that if $M \in \mathbb{R}^{m_1 \times m_2}$ and if $\Delta_{m_1} = \{\mathbf{v} \in \mathbb{R}^{m_1} : \mathbf{v} \geq 0 \sum \mathbf{v}_i = 1\}$ and similarly $\Delta_{m_2} = \{\mathbf{u} \in \mathbb{R}^{m_2} : \mathbf{u} \geq 0 \sum \mathbf{u}_i = 1\}$, then:

$$\min_{\mathbf{v} \in \Delta_{m_1}} \max_{\mathbf{u} \in \Delta_{m_2}} \mathbf{v}^\top M \mathbf{u} = \max_{\mathbf{u} \in \Delta_{m_2}} \min_{\mathbf{v} \in \Delta_{m_1}} \mathbf{v}^\top M \mathbf{u}$$

The value of the objective above, is refferred to as the value of the game M. Roughly, the theorem may be interpreted as follows: Suppose we have a matrix of payoffs M: Player 1 needs to choose a row, i, of the matrix M, and player 2 chooses a column, j, of M: Then player 1 pays player 2 the corresponding entry in the matrix (i.e. $M(i,j)$). The objective of player 1 is to minimize the cost, while player 2 wishes to maximize the cost.

The theorem states that if the players are allowed to choose a randomized strategy (i.e. each player chooses his row/column randomly according to some distribution), then the game has a well defined value, that doesn't depend on the order of the players (i.e. player 1 may choose after observing the randomized strategy picked by 2, or alternatively).

Ex. 3:

Given $S = \{\mathbf{z}^{(i)}\}_{i=1}^{m}$, we let $\sigma \in \{-1,1\}^m$ be m IID Rademacher random variables (i.e. $\sigma_i = 1$ w.p 1/2). Given a class of target function $\mathcal{F}$, denote:

$$\Re_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\mathbf{z}^{(i)})\right|\right]$$

Also denote

$$c\cdot\mathcal{F}+b = \{c\cdot f+b : f\in\mathcal{F}\}$$

$$\mathrm{conv}\mathcal{F} = \left\{\sum\lambda_i f_i : \lambda_i \geq 0, \sum\lambda_i = 1,\ f_i\in\mathcal{F}\right\}.$$

Show that:

(1) $\Re_S(c\mathcal{F}+b) = |c|\Re_S(\mathcal{F})$

(2) $\Re_S(\mathrm{conv}\mathcal{F}) = \Re_S(\mathcal{F})$.

Ex. 4:

Prove or disprove by counter example:

- If $\mathcal{F}$ is a set of convex functions then $F = \max_{f\in\mathcal{F}}(f)$ is a convex function.
- If f, g are convex functions then $f+g$ is a convex function.
- If f is a convex function and $\alpha \geq 0$ then $\alpha\cdot f$ is a convex function.
- If f, g are convex functions then $f\circ g$ is a convex function.
- If f is convex and differentiable and $\nabla f(x) = 0$ then $f(x) = \min f$.
- If for every α the set $A_\alpha = \{\mathbf{x} : f(\mathbf{x}) < \alpha\}$ is convex, then f is convex.

- If K and G are convex set then $K+G=\{\mathbf{u}+\mathbf{v}:\mathbf{u}\in K,\mathbf{v}\in G\}$ is a convex set.

Ex. 5:

A convex function f is said to be β–smooth if it is differentiable and:

$$f(\mathbf{y})\le f(\mathbf{x})+\nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})+\frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

And α strongly convex if[1]:

$$f(\mathbf{y})\ge f(\mathbf{x})+\nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})+\frac{\alpha}{2}\|\mathbf{y}-\mathbf{x}\|^2$$

- Show that for any convex loss function $\ell(\mathbf{w},(\mathbf{x},y))$, the following function $F_\lambda(\mathbf{w})$ is λ strongly convex (For simplicity, you may assume ℓ is differentiable)

$$F_\lambda(\mathbf{w})=\frac{\lambda}{2}\|\mathbf{w}\|^2+\frac{1}{m}\sum_{i=1}^m\ell(\mathbf{w},(\mathbf{x}^{(i)},y_i))$$

- Consider the loss function[2] $\ell(\mathbf{w},(\mathbf{x},y))=\log(1+\exp(-y\mathbf{w}\cdot\mathbf{x}))$, Assume $\mathbf{x}$ is bounded by $\|\mathbf{x}\|\le B$, and $y\in\{-1,1\}$. Show that ℓ is both Lipschitz bounded and convex and a smooth convex loss function (as a function of $\mathbf{w}$, for every $y,\mathbf{x}$). Calculate that parameters of Lipschitzness and smoothness.

Ex. 6:

In this exercise we will conclude that there are convex problems that are not efficiently learnable.

Let $\mathcal{H}=\{h_1,\dots,h_{2^d}\}$ be a binary hypothesis class over some domain $\mathcal{X}$. Denote $\mathbf{v}_1,\dots,\mathbf{v}_{2^d}\in\{0,1\}^d$ be all extreme points of the hyper-cube. For every i we associate $\mathbf{v}_i$ with h_i. Next define a loss

$$\ell(\mathbf{v}_i,(\mathbf{x},y))=|h_i(\mathbf{x})-y_i|,$$

and for every $\mathbf{w}\in[0,1]^d$:

$$\ell(\mathbf{w},(\mathbf{x},y))=\min\{\sum\alpha_i\ell(\mathbf{v}_i,(\mathbf{x},y)):,\ \sum\alpha_i=1,\ \alpha\ge0,\ \mathbf{w}=\sum\alpha_i\mathbf{v}_i\}$$

[1] For strong convexity we do not require the f is smooth and ∇f may stand for a subdifferential.

[2] This loss function is called the logistic loss, or logistic regression

Show that ℓ is a convex function.

(**Bonus+2**:) Show that it is also $\sqrt{d}$–Lipschitz).

(**Bonus+5**:) Conclude that there are convex learning problems that are not efficiently learnable.