Introduction to Learning (236756)
Winter 2014-15

Final exam (A)

Lecturer: Nir Ailon
Teaching assistant: Yuval Dagan

Exam rules
1. The exam contains 24 pages, including this.
2. The exam duration is 180 minutes (3 hours).
3. You may not use any material, written or electronic.
4. You can either write using a pencil or a pen, color black or blue only.
5. All answers will be written on the exam pages. All exam pages will be submitted.
6. You can use the back side of the pages for calculations. Note, however, that we provide sufficient space within the front pages.
7. All questions must be answered.
8. Use the provided spaces (inside framed boxes) for your answers.
9. When asked to choose one option among several, either mark or circle the correct option and make no mark on the other options.
10. Write legibly where asked for free text.
11. Write only what you are asked for.
12. Do not remove pages from the exam

Some good Hebrew:

<table>
<thead>
<tr>
<th>Feature</th>
<th>תכונה</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>תיבוג</td>
</tr>
<tr>
<td>More is less!</td>
<td>כל המוסיפים גורשים!</td>
</tr>
</tbody>
</table>

Good luck!
Some useful formulae

1. $\binom{n}{k} \leq n^k$

2. Generalization error = true error = $L_D^{01}$.

3. $e \approx 2.72$

4. Let $\mathcal{H}$ be a concept class of some learning problem, and $S$ a training set. Let

$$\hat{h} = \arg\min_{h \in \mathcal{H}} L_D^{01}(h)$$
$$h^* = \arg\min_{h \in \mathcal{H}} L_S^{01}(h)$$

Then for all $\delta > 0$: with probability at least $1 - \delta$:

$$L_D^{01}(\hat{h}) \leq L_D^{01}(h^*) + O\left(\sqrt{\frac{\text{VCDIM}(\mathcal{H}) + \frac{1}{\log(\delta)}}{|S|}}\right)$$
Question 1 (20 pts)

In each of the following, mark “True” or “False”. Briefly answer the corresponding textual questions when asked to do so.

1. There is known to exist an algorithm solving ERM over linear classifiers with respect to the 0/1 error in time polynomial in the problem dimension and in the number of examples.
   - True
   - False
   
   Algorithm Name:

2. The hinge loss is an upper bound on the 0/1 loss.
   - True
   - False

3. The square-loss is an upper bound on the 0/1 loss.
   - True
   - False

   Explanation:

   Counter example:
4. If $K_1$ and $K_2$ are both kernel functions, then so is $K_1 + K_2/3$

- True

Explaination:

- False

Explaination:

5. The backpropagation algorithm implements gradient descent, and hence will always lead to the optimal solution in training for any neural net over a differential activation function such as the sigmoid.

- True

- False

Explaination:

6. Naïve Bayes over binary features always outputs a linear separator.

- True

- False

Counter example:
7. K-nearest-neighbor always performs better (in terms of generalization error) as $k$ is larger (and the other parameters are left untouched)

☐ True

Explanation:


☐ False

Explanation:


8. K-nearest-neighbor always performs better (in terms of generalization error) as $k$ is smaller (and the other parameters are left untouched)

☐ True

Explanation:


☐ False

Counter example:


9. Increasing the concept class always increases the complexity of CONSISTENT for that problem

☐ True

Explanation:

☐ False

Counter example:

10. Reducing the concept class (removing hypotheses) can only reduce the size of the training set required for achieving a generalization error of $\epsilon$ using a solution for CONSISTENT (in the REALIZABLE case).

☐ True

Explanation:

☐ False

Counter example:
**Question 2 (20 pts)**

In this question $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1,1\}$. Define for all $a \in \mathbb{R}$ the following function:

$$h_a(x) = \begin{cases} +1, & x \geq a \\ -1, & x < a \end{cases}$$

Define the concept class $\mathcal{H}(2)$ as:

$$\mathcal{H}(2) = \left\{ h_{c_1,a_1,c_2,a_2,b}(x) = \begin{cases} 1, & c_1 h_{a_1}(x) + c_2 h_{a_2}(x) \geq b \\ -1, & \text{otherwise} \end{cases} : c_1, c_2, a_1, a_2, b \in \mathbb{R} \right\}$$

1. What is the VC-dimension of $\mathcal{H}(2)$? In your answer, you must illustrate a set with size $\text{VCDIM}(\mathcal{H}(2))$ and show that it can be shattered. You are encouraged to use symmetries to avoid writing too much. There is no need to find the exact values of all the parameters $c_1, a_1, ...$ but explain how to find them, and convince that we can find them. Then explain briefly why a larger set cannot be shattered.

**Solution:**
For a parameter \( k \in \mathbb{N} \) define the concept class \( \mathcal{H}_+^{(k)} \) as:

\[
\mathcal{H}_+^{(k)} = \{ h_{c_1,a_1,\ldots,c_k,a_k,b}(x) = \begin{cases} 
1, & c_1 h_{a_1}(x) + \cdots + c_k h_{a_k}(x) \geq b \\
-1, & \text{otherwise}
\end{cases} : a_1, \ldots, a_k, b \in \mathbb{R}, c_1, \ldots, c_k \geq 0 \}.
\]

Note: \( c_1 \ldots c_k \) here are nonnegative.

2. What is the VC dimension of \( \mathcal{H}_+^{(k)} \)? Prove your answer by showing a set of size \( \text{VCDIM}(\mathcal{H}_+^{(k)}) \) that can be shattered (explaining why it can be shattered), and by providing a brief explanation as to why a larger set cannot be shattered.

Answer:
Define $\mathcal{H} = \{ s \cdot h_a(x): s \in \{\pm 1\}, a \in \mathbb{R} \}$

3. For a certain classification problem, Adaboost was executed with $T$ iterations, using $a$ as a weak learner the ERM rule over $\mathcal{H}$. A hypothesis $\hat{h}$ was returned, which was not satisfying. For each of the following suggestions, determine its likely effect on the approximation error.

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Effect on approximation error</th>
<th>Effect on estimation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Increase training set</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
</tr>
<tr>
<td>B</td>
<td>Increase $T$</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
</tr>
<tr>
<td>C</td>
<td>Change the weak learner to return the ERM over $\mathcal{H}^{(2)}$</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
<td>☐ increase ☐ decrease ☐ no effect</td>
</tr>
</tbody>
</table>
Question 3 (20 pts)

1. The random variables $x_1, \ldots, x_m$ are independently distributed from the distribution $Poisson(\lambda)$. This distribution is defined for a parameter $\lambda > 0$ and assigns positive probabilities to the natural numbers \{0,1,2, ...\}. The probability of a value $x$ is given by:

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

Derive an MLE estimator for $\lambda$ (as a function of $x_1, \ldots, x_m$):

Final answer:

$$\hat{\lambda} = \left[ \right]$$

Computation:
2. A prediction problem is defined over $\mathcal{X} = \{0, 1, 2, ...\}^2$ and $\mathcal{Y} = \{0, 1\}$. It has been decided to use Naïve Bayes to learn from training data, with the assumption that for all $i = 1, 2$: $x[i]$ is distributed Poisson($\lambda_0$) conditioned on $y = 0$, and otherwise Poisson($\lambda_1$).

   a. Using the Naïve Bayes underlying assumptions, compute the following probability:

   $$\Pr[(x[1] = 5 \land x[2] = 3) \mid y = 1] = \left[ \right]$$

   b. Given the following training set of 2 labeled points $(x_1 = (0, 2), y_1 = 0), (x_2 = (2, 4), y_2 = 1)$, how would you predict the label of $x = (2, 2)$ in the Naïve Bayes framework, assuming additionally that $\Pr(y = 1) = \Pr(y = 0) = 0.5$ ?

Solution (with brief explanation):
Question 4 (20 pts)

Note: In this question you can assume that the probability of failure $\delta$ is some constant

A pediatrician is interested in predicting the onset of epilepsy by the age of 12, given the child’s medical record at age 4. Each such record contains $d$ binary features (numbers in $\{0,1\}$), which the doctor believes are relevant for such a prediction task based on her experience.

1. In the first stage, the pediatrician is trying to learn PAC using ERM. She decides on a concept class $\mathcal{H}_k = \{h: \{0,1\}^d \to \{0,1\}: h \text{ depends on } \leq k \text{ input coordinates}\}$. In words, this is the class of binary functions the output of which is determined by only $k$ input coordinates. For example, the function $x \mapsto x[1] \lor x[2] \lor x[7]$ belongs to $\mathcal{H}_3$, while the following is not in $\mathcal{H}_5$: $x \mapsto (x[3] \land x[7] \land x[9] \land x[11]) \lor (x[4] \land x[1] \land x[13])$

Write down the best upper bound on the number of samples the pediatrician would need in order to obtain a hypothesis of true error at most $\varepsilon$ above that of the optimal hypothesis in $\mathcal{H}_k$. Your bound should be expressed using $d, k, \varepsilon$. (Asymptotic expressions are ok e.g. $O(\ldots)$):

Answer: (no explanation necessary):

2. The pediatrician did not have enough labeled examples to learn using the approach in the last question, for the $k, \varepsilon$ she considered appropriate. Her resident (henceforth: resident A) proposed to give up $\frac{d}{2}$ arbitrary features. Another resident (henceforth: resident B) propose to replace $k$ with $k/2$. Which suggestion is more likely to reduce the number of required training examples more significantly?

Answer: $\square$ Resident A  $\square$ Resident B

Brief explanation:

3. Denote by $\mathcal{H}_k'$ the set of all binary valued functions on $k$ binary variables, that is, all functions from $\{0,1\}^k$ to $\{0,1\}$.

The pediatrician decided not to follow her residents’ advice, and instead to take a two phase approach.
Phase 1: A feature selection algorithm will be used to choose $k$ features (out of $d$).

Phase 2: ERM would be used to learn over $\mathcal{H}'_k$ on the training set obtained after discarding the $d - k$ features that were not chosen in phase 1.

She downloaded a software library from the internet, with a procedure called voodoo_selector. The procedure works as follows:

- $S$ a labeled set of exactly 666 sample.
- $k$ number of sought features.
- magic is a parameter in the range $\{1, 2, \ldots, k^3\}$ that affects the inner working (and the output) of voodoo_selector in a mysterious way.

The pediatrician wrote a partial program which tries all possible values of magic and chooses the best. Help her complete the program on the next page, and answer the ensuing question.

Note: Continue assuming that the pediatrician strives to obtain a generalization error which is $\varepsilon$ above that of the optimal in $\mathcal{H}_k$ (assuming voodoo_selector is as good as claimed on the website of the company that sold it to the pediatrician).
Input: $k, \varepsilon$

# Output: A set of $k$ features and a hypothesis from $\mathcal{H}_k$

# Asymptotic expressions (e.g. using $\Theta(\cdot)$) are allowed. You do not have to fill all space

$$m = \left[ \right]$$

obtain sample $S = ((x_1, y_1), \ldots, (x_{m+666}, y_{m+666}))$

for $\text{magic} = 1 \ldots k^3$:

$$(f[1], \ldots, f[k]) = \text{voodoo_selector}((x_{m+1}, y_{m+1}), \ldots, (x_{m+666}, y_{m+666}), k, \text{magic})$$

define $x'_i = (x_i[f[1]], \ldots, x_i[f[k]])$ for all $i = 1 \ldots m$

$$h = \text{ERM} \left( \mathcal{H}'_k, \left( \left( x'_1 \right), y_1 \right), \ldots, \left( x'_m \right), y_m \right)$$

$$\text{err} = L_{01} \left( h, \left( x'_1 \right), y_1 \right), \ldots, \left( x'_m \right), y_m \right)$$

end for

return $\left[ \right]$
Provide a brief justification for the initialization of \( m \) in the code:
Question 5 (20 pts)

Yossi has two binary classification problems. For both problems $\mathcal{X} = \mathbb{R}^2$.

For each problem, he randomly collected samples.

The first problem is related to sandwich preferences.

The labeled examples are as follows:

![Graph of sandwich preferences](image1)

The second is related to course preferences. The labeled examples are as follows:

![Graph of course preferences](image2)
Recall the following definitions:

\[ L_{S}^{\text{hinge}}(w) = \frac{1}{|S|} \sum_{(x,y) \in S} \max\{0, 1 - \langle w, x \rangle y \} \]

\[ L_{S,\lambda}^{\text{hinge}}(w) = \frac{\lambda}{2} \|w\|^2 + L_{S}^{\text{hinge}} \]

\[ L_{S}^{01}(w) = \frac{1}{|S|} \sum_{(x,y) \in S} \begin{cases} 1, & \text{sign}(\langle w, x \rangle) \neq y \\ 0, & \text{otherwise} \end{cases} \]

For each of the two problems, Yossi divided the labeled example set into two equal sized sets: a training set \( S_{\text{train}} \) and a test set \( S_{\text{test}} \). He then tried to find a coefficient vector \( w \in \mathbb{R}^2 \) minimizing \( L_{S_{\text{train}},\lambda}^{\text{hinge}} \).

To that end, he used sub-gradient descent, with step size \( \eta \). More precisely, he executed the following:

- \( w^{(0)} \leftarrow \text{random} \in \mathbb{R}^2 \)
- \( \text{for } t = 1, \ldots, 250 \)
  - \( w^{(t)} \leftarrow w^{(t-1)} - \eta \nabla L_{S_{\text{train}},\lambda}(w^{(t-1)}) \)

For each of the two classification problems (sandwiches, courses) he made 4 executions, with the following corresponding parameter values:

- \( \lambda = 0, \eta = 0.1 \)
- \( \lambda = 0.5, \eta = 0.1 \)
- \( \lambda = 0, \eta = 0.5 \)
- \( \lambda = 0.5, \eta = 0.5 \)

Yossi printed two reports, one for the sandwiches and the other for the courses. It is not known which report corresponds to which problem.
First report:

Second report:

A

B

C
Notice that each report contains 3 diagrams.

- One diagram shows $L_{\text{Strain}}^{\text{hinge}}(w(t))$ as a function of $t$
- One diagram shows $L_{\text{Test}}^{\text{hinge}}(w(t))$ as a function of $t$
- One diagram shows $L_{\text{Strain}}^{01}(w(t))$ as a function of $t$

Each diagram includes 4 plots (blue lines), corresponding to the 4 settings of the parameters $\lambda, \eta$ as detailed above. The number $r'$ appearing on the $y$ axis of each graph is some common positive constant.

1. Which problem (sandwiches or courses) does the first report correspond to? Provide a brief explanation.

- Sandwiches
- Courses

Short explanation
2. Observe the second report. Find the correct correspondence between the diagrams in the report, and
the loss functions. Provide a brief explanation.

\[ L_{\text{train}}^{\text{hinge}}(w(t)) \] corresponds to diagram  ☐ A  ☐ B  ☐ C

\[ L_{\text{test}}^{\text{hinge}}(w(t)) \] corresponds to diagram  ☐ A  ☐ B  ☐ C

\[ L_{\text{train}}^{01}(w(t)) \] corresponds to diagram  ☐ A  ☐ B  ☐ C

Brief explanation:
3. Continuing with the second report, diagram B is enlarged here for your convenience:

The plots are marked with Hebrew letters \( \tau, \lambda, \beta, \kappa \). Find the correct correspondence between the parameter values and the plots. (No need to explain)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>( \lambda )</td>
<td>( \beta )</td>
<td>( \kappa )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda = 0, \eta = 0.1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda = 0, \eta = 0.5 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda = 0.5, \eta = 0.1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda = 0.5, \eta = 0.5 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. An additional diagram for the second report was found:

The diagram shows the norm of \( w(t) \) as a function of \( t \). Find the correspondence between the parameter values and the 4 plots in the diagram.

\[
\begin{array}{cccc}
\lambda & \eta \\
0 & 0.1 \\
0.5 & 0.1 \\
0 & 0.5 \\
0.5 & 0.5 \\
\end{array}
\]