Technion – Israel Institute of Technology ***Summer 2017***

Electrical Engineering

Introduction to Machine Learning – International School

Final Exam

1. Exam duration is **three hours**.
2. It is highly recommended to read the entire exam before you start.
3. Include explanations. You should answer all questions. The value of each question is given in the body of the question. The total number of points is 100.
4. You may use any material during the exam (including laptops).
   Disable any connectivity.
5. Write in English, in a clear and organized manner.
6. The exam sheet includes 6 pages, including this page.

Good Luck! 祝你好运!

Some useful formulas:

$\sum_{k=0}^{k=\infty} q^k = \frac{1}{1-q}, \text{for } 0 < q < 1$
$\sum_{k=1}^{k=\infty} kq^{k-1} = \frac{1}{(1-q)^2}, \text{for } 0 < q < 1$
$\sum_{k=0}^{k=\infty} \frac{q^k}{k!} = e^q, \text{for } q \in \mathbb{R}$
$\sum_{k=1}^{k=\infty} \frac{q^k}{k} = -\log(1-q), \text{for -1} \le q < 1$

Question 1 – Bayesian Classification (37%)

We examine the problem of "spam" document identification using a simple "bag-of-words" model. Given a dictionary with J words, denoted by $\{1,...,J\}$, we denote:

- $p_0 \in (0,1)$ is the relative frequency of spam documents.
- $\vec{m} = (m_1,...,m_J)^T$ is the vector of word frequencies, where m_j is the number of times the word j has appeared in a given document.
- $y \in \{-1,1\}$ is the document label, where $y=-1$ denotes "spam".
- $P(m_j = m \mid y=-1) = A_j \cdot (\alpha_j)^m$ is the probability that a word j would appear m times in the document, given that it is spam. (A_j, α_j) are given numbers, where $\alpha_j \in (0,1)$.
- $P(m_j = m \mid y=1) = B_j \cdot (\beta_j)^m$ is the probability that a word j would appear m times in the document, given that it is not spam. (B_j, β_j) are given numbers, where $\beta_j \in (0,1)$.

We want to use the Naïve Bayes model which classifies the document based on the word-frequency vector $\vec{m}$, and the probabilistic model we defined. Assume that all constants are known, unless stated otherwise. The risk function is the standard error probability (0-1 loss).

1. (9%) Write the probability that a document with frequency vector $\vec{m}$ is spam (given the assumptions of the model above), as a function of $\alpha_j, \beta_j, A_j, B_j, p_0, J$ and the components of $\vec{m}$.
2. (8%) Prove that the optimal decision rule resulting from the "Naïve Bayes" assumption is linear, of the form $\hat{y} = \text{sign}\left(w^T \vec{m} + b\right)$, and find w, b explicitly, as a function of $\alpha_j, \beta_j, A_j, B_j, p_0$, and J.
3. (a) (4%) Find the constants A_j, B_j as a function of α_j and β_j.

   (b) (8%) In this part of the question (only) we want to estimate the constants α_j from examples. You are given N spam documents, with frequency vectors $\{\vec{m}(i)\}_{i=1}^{N}$. Find the maximum likelihood (MLE) estimate of each parameter α_j, as a function of N and the components of $\vec{m}$.

In addition to the given classifier, a linear classifier was trained using ERM (empiric risk minimization) using the 0-1 loss on $N < \infty$ documents and the result was a decision rule of the form $\hat{y}_{ERM} = \text{sign}\left(\tilde{w}^T \vec{m} + \tilde{b}\right)$.

4. (a) (3%) Write explicitly the empiric risk that should be minimized. Assume from now on that the optimal parameters $\left(\tilde{w},\tilde{b}\right)$, which minimize this risk, can be found.

(b) (5%) Assume that $\tilde{w}$ is a unit vector, i.e. $\|\tilde{w}\|=1$. Are the optimal parameters necessarily unique? Explain and demonstrate graphically in the 2D case.

Solution

1. From Bayes Rule

$$P\left(y=-1\,|\,\vec{m}\right)=\frac{P\left(\vec{m}\,|\,y=-1\right)P\left(y=-1\right)}{P\left(\vec{m}\,|\,y=-1\right)P\left(y=-1\right)+P\left(\vec{m}\,|\,y=1\right)P\left(y=1\right)}$$

Also, from the Naïve Bayes assumption:

$$P\left(\vec{m}\,|\,y=-1\right)=\prod_{j=1}^{J}P\left(m_j\,|\,y=-1\right)=\prod_{j=1}^{J}A_j\cdot\alpha_j^{m_j}$$

$$P\left(\vec{m}\,|\,y=1\right)=\prod_{j=1}^{J}P\left(m_j\,|\,y=1\right)=\prod_{j=1}^{J}B_j\cdot\beta_j^{m_j}$$

Therefore

$$P\left(y=-1\,|\,\vec{m}\right)=\frac{p_0\prod_{j=1}^{J}A_j\cdot\alpha_j^{m_j}}{p_0\prod_{j=1}^{J}A_j\cdot\alpha_j^{m_j}+\left(1-p_0\right)\prod_{j=1}^{J}B_j\cdot\beta_j^{m_j}}$$

2. We decide the $y=1$ if

$$\log P\left(y=1\,|\,\vec{m}\right)>\log P\left(y=-1\,|\,\vec{m}\right)$$

$$\Rightarrow\log\left(1-p_0\right)+\sum_{j=1}^{J}\log B_j+m_j\log\beta_j>\log p_0+\sum_{j=1}^{J}\log A_j+m_j\log\alpha_j$$

$$\Rightarrow\log\left(\frac{1-p_0}{p_0}\right)+\sum_{j=1}^{J}\log\frac{B_j}{A_j}+\sum_{j=1}^{J}m_j\log\frac{\beta_j}{\alpha_j}>0$$

$$\Rightarrow b=\log\left(\frac{1-p_0}{p_0}\right)+\sum_{j=1}^{J}\log\frac{B_j}{A_j},w_j=\log\frac{\beta_j}{\alpha_j}$$

3. (a) From normalization, the probabilities sum to 1. Therefore

$$1=\sum_{m=0}^{\infty}A_j\alpha^m=A_j\sum_{m=0}^{\infty}\alpha_j^m=A_j\frac{1}{1-\alpha_j}$$

Which implies that $A_j=1-\alpha_j$. Similarly, $B_j=1-\beta_j$.

(b) We write the likelihood

$$L=P\left(\left\{\vec{m}(i)\right\}_{i=1}^{N} \mid \left\{y(i)=-1\right\}_{i=1}^{N}\right)=\prod_{i=1}^{N}P\left(\vec{m}(i)\mid y(i)=-1\right)$$

$$=\prod_{i=1}^{N}\prod_{j=1}^{J}A_j\cdot\alpha_j^{m_j(i)}=\prod_{j=1}^{J}\left(1-\alpha_j\right)^{N}\alpha_j^{\sum_{i=1}^{N}m_j(i)}$$

Denoting $M_j=\sum_{i=1}^{N}m_j(i)$, we differentiate the log likelihood by α_j

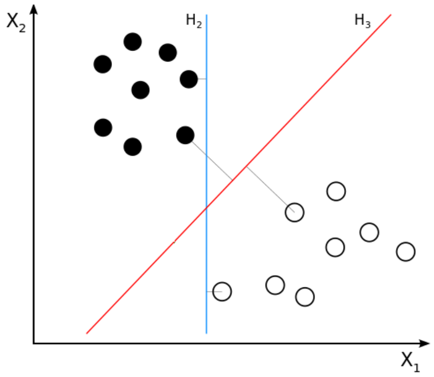
$$0=\frac{\partial}{\partial\alpha_j}LL=\frac{\partial}{\partial\alpha_j}\left[N\log\left(1-\alpha_j\right)+M_j\log\alpha_j\right]=\frac{-N}{1-\alpha_j}+\frac{M_j}{\alpha_j}$$

$$\Rightarrow\frac{N}{1-\alpha_j}=\frac{M_j}{\alpha_j}\Rightarrow\alpha_j=M_j\left(1-\alpha_j\right)\Rightarrow\frac{M_j}{N+M_j}=\alpha_j$$

4. (a) The empiric risk with the 0-1 loss is:

$$\hat{R}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{I}\left[y(i)\neq\operatorname{sign}\left(\tilde{w}\vec{m}(i)+b\right)\right]$$

(b) No, the optimal parameters are not unique, since one can pass an infinite number of lines which give the same classification. For example, in the linearly separable case:

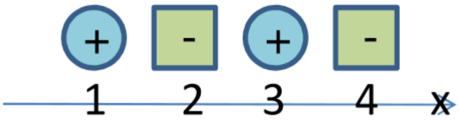


But this is also true in the non-separable case.

Question 2 – Decision Trees, Boosting (39%)

Note: This question contains 3 parts which are unrelated

1. (12%) Consider the following training set of 4 samples in 1 dimension (see figure). We define the set of functions $H=\{\text{sign}(x-b):b\in\mathbb{R}\}$, where $\text{sign}(x)=1$ if $x\geq 0$ and $\text{sign}(x)=-1$ if $x<0$.



   a) (4%) A decision tree algorithm is applied to this training set, with decision functions $h\in H$ in each node. Is there such a tree with 0 error on the training set? Explain. If the answer is yes, find such a tree with minimal number of nodes.

   b) (4%) Are there classifiers $\{h_t\in H\}$ and parameters $\{\alpha_t\}$ such that a hypothesis of the form $\text{sign}\left(\sum_t \alpha_t h_t(x)\right)$ achieves 0 error on the training set? Explain. **Note:** the parameters α_t are not constrained to be positive.

   c) (4%) Repeat (a) where instead of H, you use $G=\{\text{sign}(ax^2+bx+c):a,b,c\in\mathbb{R}\}$.

2. (12%) The following parts are unrelated.

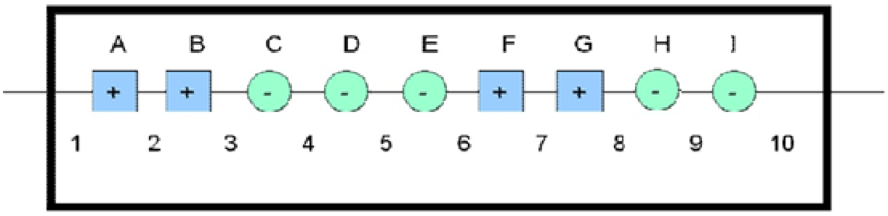
   a. (4%) Determine if the following statement is correct, explain briefly: Two different decision trees that label the same training set identically with training error zero, will label every input identically.

   b. (4%) Assume a learning problem that has many noisy features (i.e. features with no correlation with the label). Which algorithm is better in this case, decision tree or 1-NN. Explain. (A qualitative explanation is enough).

   c. (4%) For this question, we say that 2 trees are the same if they perform the same splits in all nodes on a given training set. Assume a classification problem with a training set over $\Re^{100}$. Student A trained a decision tree (as we learned in class). Student B first normalized the training set so the average of every feature is 0 and the standard deviation is 1. Then he trained the exact same algorithm as student A. Which of the following is correct:

      i. Both students will necessarily get the same tree.
      ii. Both students will necessarily get different trees
      iii. At times, student A and Student B will receive the same tree, and at times, they will receive different trees.

3. (15%) The AdaBoost algorithm is applied to the training set sketched below (the samples are located on the real line). In every iteration the weak learner h_t is a threshold function in one of the ten samples numbered 1-10. The labeling can be positive or negative for every side, meaning that there are total of 20 options for selecting a threshold function. We denote by "A" the algorithm that selects h_t in every step such the error ϵ_t is minimal.

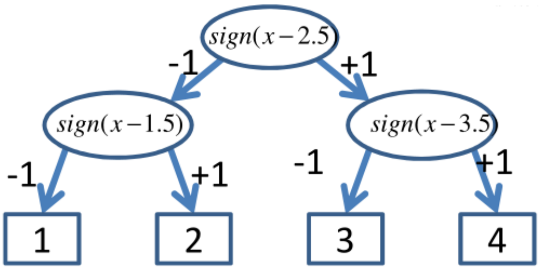


   a. (5%) Which weak learner will the algorithm A choose in the first iteration? (Write down the corresponding threshold value, and state whether the samples on its right will be labeled with "+" or "–").
   b. (5%) Which weak learner will the algorithm A choose in the second iteration? (write down the corresponding threshold value, and state whether the dots on its right will be labeled with "+" or "–").
   c. (5%) Assume that the AdaBoost algorithm stopped after the above two steps. Does the final classification function has zero error on the training set?

Question 2 – Solution

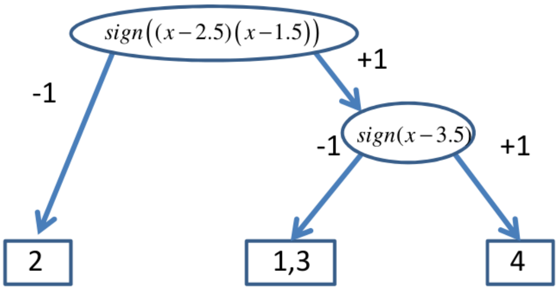
1.

a. There is a tree with 0 training error:



b. There is such function, for example : $f(x) = -\text{sign}(x-1.5) + \text{sign}(x-2.5) - \text{sign}(x-3.5)$. Note that we must use negative coefficients, otherwise all weak learners will have error greater than 0.5 already in the 1st iteration.

c. There is a smaller tree:



2.

a. The statement is not true, for example the 2 trees may have in some node a threshold of 0.9 in one tree and 1.1 in the other tree. If all training examples have in the corresponding feature values which are either smaller than 0.9 or larger than 1.1, these nodes will split the training set in the same manner. However, a new example with value of 1.0 in that feature, will be sent to different branches.

b. A decision tree is better in that case, since it can simply ignore the noisy features by not using them in any node, while the 1-NN algorithm cannot ignore features.

c. The correct answer is (i). Normalizing to 0 mean and standard deviation of 1 is done by transforming x to $\frac{x-\mu}{\sigma}$. The algorithm can choose at each node a `transformed' threshold $\frac{t^{*}-\mu}{\sigma}$ instead of t^{*}, resulting in the same tree.

3.

a. At iteration 1 the threshold will be 3, labeling '+' to the left. Samples F, G will be wrongly labeled, thus the error is $\epsilon_1 = \frac{2}{9}$ giving $\alpha_1 = \frac{1}{2}\log\frac{7}{2}$.

b. At iteration 2, we get $D_2 = \frac{1}{28}[2,2,2,2,2,7,7,2,2]$, so the algorithm will choose the threshold at 8, labeling '+' to the left. Samples C, D, E will be wrongly labeled, so $\epsilon_2 = \frac{3}{14}$ giving $\alpha_2 = \frac{1}{2}\log\frac{11}{3}$.

c. The prediction of the final hypothesis on the training set is:

$$\begin{aligned} H &= \text{sign}\left(\frac{1}{2}\log\frac{7}{2}[+,+,-,-,-,-,-,-,-] + \frac{1}{2}\log\frac{11}{3}[+,+,+,+,+,+,+,-,-]\right) \\ &= \text{sign}\left([1.27, 1.27, 0.02, 0.02, 0.02, 0.02, 0.02, -1.27, -1.27]\right) \\ &= [+,+,+,+,+,+,+,-,-] \end{aligned}$$

so the algorithm does not achieve 0 training error.

<u>Question 3 - Misc. Topics (24%)</u> Note: This question has 3 sub-questions which are unrelated.

1. (8%) Assume $x \in \mathbb{R}$. Consider the transformation $\phi(x) = (\phi_0(x), \phi_1(x), \ldots)^T$ where

$$\phi_m(\text{x}) = \frac{x^m}{\sqrt{m!}} e^{-\frac{x^2}{2}}, \quad m = 0, 1, 2, \ldots$$

Note that the transformed vector $\phi(x)$ is of infinite dimension. Calculate the kernel function $k(x,z)$ associated with ϕ. Show that it can be written in the form $k(x,z) = f(x-z)$ for some function f, and calculate f explicitly.

2. (8%) The perceptron algorithm for a set of samples $\{x_t, y_t\}_{t=0}^{n-1}$, where $y_t \in \{-1,1\}$, can be defined as follows:

- $w_0 = 0$
- for $t = 0,1,2,...,n-1$: $\hat{y}_t = \text{sign}\left(w_t^T x_t\right)$

$$w_{t+1} = w_t + \frac{1}{2}(y_t - \hat{y}_t)x_t$$

We apply transformation $\phi(\cdot)$ to the samples such that x_t in the algorithm is replaced by $\phi(x_t)$. Since the dimension of $\phi(x_t)$ can be very high (or even infinite), we need to use a kernel $k(x,z)$ which is given for the transformation $\phi(x_t)$, that is assume that for any two samples x,z the inner product is known: $\phi(x)^\top \phi(z) = k(x,z)$.

Propose a way to apply the perceptron algorithm for this case, using the given kernel and without using the transformation $\phi(x)$ explicitly.

3. (8%) For a specific SVM problem with 2 samples x_1, x_2 (and the regular inner product), we have the following dual problem:

$$\max_{\alpha_1,\alpha_2}\{\alpha_1 + \alpha_2 - \tfrac{1}{2}(\alpha_1)^2 - 2(\alpha_2)^2 + L\alpha_1\alpha_2\}$$
$$s.t.: \quad \alpha_1 \geq 0, \quad \alpha_2 \geq 0, \quad \alpha_1 - \alpha_2 = 0$$

where L is a constant number.

Out of the following values:

$$-3, -1, 0, 1, 3$$

what are the possible values for the constant L? Explain your answer.

Hint: Recall the Cauchy-Shwartz inequality $|\mathbf{x}\cdot\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$.

Solution

1.

$$\langle\phi(\text{x}),\phi(\text{z})\rangle = \sum_{n=0}^{\infty}\left(\frac{x^n}{\sqrt{n!}}e^{-\frac{x^2}{2}}\right)\left(\frac{z^n}{\sqrt{n!}}e^{-\frac{z^2}{2}}\right) =$$

$$e^{-\frac{x^2+z^2}{2}}\sum_{n=0}^{\infty}\frac{(xz)^n}{n!} = e^{-\frac{x^2+z^2}{2}}e^{-(\text{xz})} = e^{-\frac{x^2-2xz+z^2}{2}} = e^{-\frac{\|x-z\|^2}{2}}$$

2. The objective function of the dual problem is:

$$\max \alpha_1 + \alpha_1 - \frac{1}{2}\alpha_1{}^2 \|x_1\|^2 - \frac{1}{2}\alpha_1{}^2 \|x_1\|^2 + \alpha_1\alpha_2 x_1^T x_2$$

By comparison to the objective given in the question, we get:

$$\|x_1\|^2 = 1$$
$$x_1^T x_2 = L$$
$$-\frac{1}{2}\|x_2\| = -2 \rightarrow \|x_2\|^2 = 4$$

Using the last result and the Cauchy-Schwartz inequality we get

$$L = x_1^T x_2 = < x_1, x_2 >$$
$$\downarrow$$
$$-\|x_1\| \cdot \|x_2\| \le |< x_1, x_2 >| \le \|x_1\| \cdot \|x_2\|$$

Summarizing, we get the condition for L :

$$-2 = -\|x_1\| \cdot \|x_2\| \le L \le \|x_1\| \cdot \|x_2\| = 2$$

So the possible values are $-1, 0, 1$.

3. Since the weight vector is initialized to zeros, it is easy to see that

$$w_t = \sum_{i \in I_t} y_i \phi(\mathrm{x}_i)$$

where I_t is the set of indices of samples on which the algorithm made mistakes up to step t.

The algorithm therefor takes the form:

```
input : $w_0 = (0, 0, \ldots)$
for $t = 1, 2, \ldots$
        $\hat{y}_t = \sum_{i \in I_t} y_i \phi(\mathrm{x}_i)^T \phi(\mathrm{x}_t)$
        $w_{t+1} = w_t + \frac{1}{2}(\mathrm{y}_t - \hat{\mathrm{y}}_t)\phi(\mathrm{x}_t)$
```

The algorithm can be written even more compactly as:

```
$\text{input} : I = \{\}$
$\text{for } t = 1,2,\ldots$
    $\hat{y}_t = \sum_{i \in I} y_i \phi(\mathrm{x}_i)^T \phi(\mathrm{x}_t)$
    $\text{if } \operatorname{sign}(\mathrm{y}) \neq \operatorname{sign}(\hat{y}_t)$
        $I = I \cup \{\mathrm{t}\}$
```

And of course, now the algorithm depends only on inner products, so we do not need the transformation explicitly, and we can use the kernel function:

```
$\text{input} : I = \{\}$
$\text{for } t = 1,2,\ldots$
    $\hat{y}_t = \sum_{i \in I} y_i K(\mathrm{x}_i, \mathrm{x}_t)$
    $\text{if } \operatorname{sign}(\mathrm{y}) \neq \operatorname{sign}(\hat{y}_t)$
        $I = I \cup \{\mathrm{t}\}$
```