Introduction to Machine Learning (67577)

Problem Set 1 – Linear Algebra and Probability Review

Due: 23.3.2016

Submission guidelines

- The assignment is to be submitted via Moodle only.
- Files should be zipped to ex_N_First_Last.zip (In this case N stands for 1, First is your first name, Last is your last name. In English, of course).
- The .zip file size should not exceed 10Mb.
- All answers to both theoretical part and practical part, should be submitted as a single pdf file. We encourage you to typeset your answers (either Lyx or Word), but it is not mandatory. Students who choose to submit handwritten scans are warned that incomprehensible answers due to handwriting or bad scan, will result in point reduction just like wrong answers are. Moreover, the scans should be integrated with the plots to a single coherent pdf file.
- All of your Python code files used in the practical question should be attached in the zip file. There should be one .py file, other then that there are no special requirements or guidelines for your code. Note that if your code doesn't run on the CS environment you will not get points for this question.

1 Thresholds: lower bounds

In the context of online learning, let's define 3 notions:

- "a realizable environment" means an environment that follows some $h^* \in \mathcal{H}$, i.e for all t, $y_t = h^*(x_t)$, where x_t is the sample that the environment produces in time t and y_t is its label.
- $M_{\mathcal{A}}(E)$ is the number of mistakes the $\mathcal{A}$ does when E produces the samples and their labels.

- $M_{\mathcal{A}}(\mathcal{H})$ is the maximal number of mistakes $\mathcal{A}$ might make on a sequence of examples which is labeled by a realizable environment, i.e $M_{\mathcal{A}}(\mathcal{H}) = max_E \{M_{\mathcal{A}}(E)\}$. When $\mathcal{H}$ is clear from the context we may simply write $M_{\mathcal{A}}$

Let $N \geq 2$. Consider the online prediction problem whose domain is the set $\mathcal{X} = [N-1] = \{1, ..., N-1\}$, its label set is the set $\mathcal{Y} = \{-1, 1\}$, the hypothesis class $\mathcal{H}$ is the class of thresholds over $\mathcal{X}$, that is,

$$\mathcal{H} = \left\{ h_\theta(x) = sign(x - \theta) : \theta \in \left\{ \frac{1}{2}, \frac{3}{2}, ..., N - 1 + \frac{1}{2} \right\} \right\}$$

and the number of rounds n is at least $\lfloor \log N \rfloor$. Let A be a learning algorithm.

1. (10 points) Show that there exists a realizable environment E which satisfies $M_{\mathcal{A}}(E) \geq \lfloor \log |\mathcal{H}| \rfloor = \lfloor \log N \rfloor$.

2. (10 points) Replace the set $\mathcal{X}$ in the previous part with the set $[0, 1]$ (now θ ranges over $[0, 1] \cup \left\{-\frac{1}{2}, \frac{3}{2}\right\}$). Prove that for every learning algorithm $\mathcal{A}$, and any number of rounds n, $M_{\mathcal{A}}(\mathcal{H}) = n$.

2 Linear algebra - SVD

1. (9 points) In this question we will show the relationship between the singular value decomposition and the left and right singular vectors. Let $A \in \mathbb{R}^{m \times d}$ with SVD $A = U \Sigma V^\top$. Denote: v_i, u_i the ith columns of V, U respectively and $\sigma_i = \Sigma_{ii}$. Show: $Av_i = \sigma_i u_i$ and $A^\top u_i = \sigma_i v_i$.

2. (2 points) Explain why we may assume $\sigma_i \geq 0$ for all $i \in \{1, 2, ..., \min\{m, d\}\}$.

3. (9 points) Prove the relation between SVD and EVD mentioned in tirgul1. let $A \in \mathbb{R}^{m \times d}$. Prove:

   (a) If $A = U \Sigma V^\top$ is an SVD of A, then $AA^\top = U \Sigma \Sigma^\top U^\top$ is an EVD of $AA^\top$ and $A^\top A = V \Sigma^\top \Sigma V^\top$ is an EVD of $A^\top A$.

   (b) Denote $rank(A) = k \leq \min(m, d)$. If $A^\top A = V D V^\top$ is an EVD of $A^\top A$ then we can find an orthonrmal basis for $\mathbb{R}^m$, denoted by $\overline{u}_1, \overline{u}_2, ..., \overline{u}_m$ such that $A = U \Sigma V^\top$ is an SVD of A and $AA^\top = U \Sigma \Sigma^\top U^\top$ is an EVD of $AA^\top$. Where: $\mathbb{R}^{m \times m} \ni U = (\overline{u}_1, \overline{u}_2, ..., \overline{u}_m)$ and $\Sigma \in \mathbb{R}^{m \times d}$ satisfies that $\Sigma^\top \Sigma = D$. Hint: denote $\overline{v}_i$ as the ith column of V and define $\overline{w}_i = A\overline{v}_i$. Explain

why the non-zero vectors among $\{\overline{w}_1, \overline{w}_2, ..., \overline{w}_n\}$ (how many are there?) are orthogonal. Normalize the $\overline{w}_i$s and complete them to an orthonormal basis for $\mathbb{R}^m$.

3 Projection matrices

Let V be a k-dimensional subspace of $\mathbb{R}^d$, and let $v_1, v_2, ..., v_k \in \mathbb{R}^d$ be an orthonormal basis of V . Define $P = \sum_{i=1}^{k} v_i v_i^\top$. The matrix P is called a "projection matrix" onto the subspace V. Prove the following useful properties of projection matrices:

1. (4.5 points) P is symmetric.
2. (4.5 points) The vectors $v_1, v_2, ..., v_k$ are eigenvectors of P, all with eigenvalues equal to 1.
3. (1 point) Write a possible EVD of P.
4. (4.5 points) $P^2 = P^\top P = PP^\top = P$
5. (1 point) $(I - P)\, P$ is the 0 matrix.
6. (4.5 points) $x \in V \Rightarrow Px = x$.

4 Linear algebra - norms and affine transformations (bonus question)

1. (1 point) Define and draw the unit ball in $\mathbb{R}^2$ under the L_1 norm, $\mathcal{B}_1 = \left\{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\right\}$. In you definition use "coordinate language" (e.g., $\left\{x \in \mathbb{R}^2 : x_1 = 17 \text{ and } \left(x_2^2 \leq 4 \text{ or } x_2 > 6\right)\right\}$).
2. (1 point) Define (in coordinate language) and draw the unit ball in $\mathbb{R}^2$ under the L_∞ norm, $\mathcal{B}_\infty = \left\{x \in \mathbb{R}^2 : \|x\|_\infty \leq 1\right\}$.
3. (3 points) Find a linear transformation M such that $\mathcal{B}_\infty = \{Mx : x \in \mathcal{B}_1\}$
   - To find M, write a system of 2 linear equations with 4 variables that maps the points $(1, 0), (0, 1)$ in $\mathcal{B}_1$ to the points $(1, 1), (-1, 1)$ in $\mathcal{B}_\infty$ and find its unique solution.
   - Prove that if $x \in \mathcal{B}_1$ then $Mx \in \mathcal{B}_\infty$, you are not required to prove the other direction.

5 Concentration inequalities - a practical question

In this practical question we visualize the concentration inequalities seen in tirgul 2. Reminder: we saw 2 results regarding the sample complexity (the number of samples needed to ensure the approximation (ϵ) and confidence (δ) levels desired). The first result is corollary 2 (that uses Chebyshev's inequality) and the second result is corollary 4 (that uses Hoeffding's inequality). We would like to compare between these 2 bounds with empirical data.

1. (5 points) We have a coin with unknown bias p that we wish to estimate with accuracy ϵ and confidence δ. What are the m we found in tirgul 2 such that if we have more than m i.i.d samples of our coin we can estimate p with the accuracy and confidence needed.

2. (35 points) You are given 100000 sequences of 1000 coin tosses (arranged in a matrix, "data", of 100000 rows and 1000 columns. To generate the data use the commands:

   - import numpy
   - data = numpy.random.binomial(1, 0.25, (100000,1000))

   Define a variable "epsilon" which gets the values $[0.5, 0.25, 0.1, 0.01, 0.001]$. For each ϵ:

   (a) For the first 5 sequences of 1000 tosses (the first 5 rows in "data"), plot the estimate $\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} x_i$ as a function of m (i.e the mean of all tosses up to m). 1 figure with 5 plots (each row in a different color). What do you expect to see in this plot as m grows?

   (b) For each bound (Chebyshev and Hoeffding) and for each ϵ, plot the upper bound on $P_{X_1,\ldots,X_m}\left(\left|\overline{X}_m - \mathbb{E}[X]\right| \geq \epsilon\right)$ (derived in class) as a function of m (where m ranges from 1 to 1000). 5 figures with 2 plots each (mention in the title of each plot what is ϵ and use a different color for each bound)[1].

   (c) You are now told that $p = 0.25$. On top of the figures from the previous question, plot the percentage of sequences that satisfy $\left|\overline{X}_m - \mathbb{E}[X]\right| \geq \epsilon$ as a function of m (now you know $\mathbb{E}[X] = p = 0.25$). What are you expecting to see in these plots? Explain.

[1] if the bound calculated is greater than 1, you should substitute it with 1. This is because we are bounding a probability so a bound greater then 1 is irrelevant.