In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 1

Spring Semester 1402-03

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. R. Amiri

Due on Esfand 20, 1402 at 23:55

(∗) starred problems are optional!

1 All You Need is Combinatorics

In this problem, you are given an *Artificial Machine*(!) and you are asked to teach it in a way that in the end, it would *Learn Intelligence*(?). So basically, by solving this problem completely, you will know *everything* needed in this course.

1.1 A simple Random Walk

First, suppose that our Machine shows an integer at each time and it starts with 0 at the first round. On each round after that, the Machine randomly and with equal probability, either increases its number by 1 or decrease it by 1 (it can show negative integers too). Calculate the probability of our Machine showing 0 on its monitor again, if we know that it will work for exactly T rounds. (There is no need for your answer to have a closed and simple form and just the solution is important for this part)

1.2 Everything is possible!

According to the previous part, show that if T goes to ∞, then the probability of seeing 0 again will converge to 1. Also with a similar reasoning show that we will see every other number at least once too(for this, no calculations is needed).

1.3 A naive algorithm

Now, we want to teach the first *Learning* algorithm to our Machine. For this, suppose that we have n secrets which are either 0 or 1 with equal probability. In other words, our whole secret is a random vector of 0 and 1's in $\{0,1\}^n$, which the Machine isn't aware of. At each round, Machine guesses a random vector of the same form totally random(with same probability for each vector), and it will stop whenever its guess completely match the secret. Calculate the expected time of success for our algorithm and based on that, explain why in practice, using such an algorithm isn't helpful at all.

1.4 A less naive algorithm

Now, suppose that our hidden secret is a *Random Permutation* of $1, 2, \ldots, n$. At each round, Machine guesses the first unrevealed position of this permutation and when it guess becomes true, it will move to the next position. Algorithm will end whenever Machine guesses all of the secret. The algorithm which Machine makes guess at each round is as follows: for each position of permutation, Machine start guessing from the least unrevealed number, and after each wrong guess, guesses the next least unrevealed number. Prove that this algorithm has the best performance, in the means of expected number of time for success. Also, calculate the expected number of times this Machine will guess before success.

1.5 Gaussians everywhere

In this part, assume that our secret is a random vector from *Normal*, n-dimensional distribution. It means that each of it's entries are from Normal distribution(Gaussian with Mean 0 and Variance 1). Now, our algorithm is upgraded so much and it works like this: It starts from an arbitrary vector. At each round, it randomly selects an index. After that, it samples a random number from Normal distribution like x. Then it either adds or removes x from the selected index, and it chooses the one which is closer to the secret in that index(assume that somehow we can know which one is closer). The algorithm will stop whenever at each index, difference between algorithm's number and secret, is less than some fixed t. First show that the best starting point for the algorithm is 0^n vector, then calculate the expected finishing time of algorithm(there is no need to explicit calculation for the expected time and only a sketch of the proof suffices).

2 Statistics and other friends

2.1 LLN & CLT

Briefly explain *Law of Large Numbers* and *Central Limit Theorem*.

2.2 Assumption

Discuss how this two theorems and their implications are related to *Statistics*. How do you think they are going to be used in the course?

2.3 Are you sure about that?

Briefly explain *Hypothesis Test* and *Confidence Interval*.

2.4 Another Assumption

Discuss how this two concepts and their implications are related to *Statistics*. How do you think they are going to be used in the course?

2.5 Time to take out your pen!

Consider $X_1, X_2, \ldots, X_n$ as n independent random variables, having the same distribution as random variable X from $[0, 1]$ interval. Also consider $Y_1, Y_2, \ldots, Y_n$ as Bernouli random variables independent from each other and also independent from $X_1, X_2, \ldots, X_n$, each with parameter $X_1, X_2, \ldots, X_n$ respectively. You are given the values of $Y_1, Y_2, \ldots, Y_n$, also we don't know the values of $X_1, X_2, \ldots, X_n$. Based on this, create a 95% confidence interval for $\mu = \mathbb{E}[X]$.

3 Its all about Tails

3.1 Not a very hard inequality

Consider random variable X, with $\mathbb{E}[X] = 0$ and $Var[X] = \sigma^2$. show that for each $a > 0$, we have:

$$\mathbb{P}[X \geq a] \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

3.2 Gaussian

Show that for $X \sim \mathcal{N}(0, \sigma^2)$ and for each $s \in \mathbb{R}$, we have:

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}.$$

3.3 Under the Gaussian!

Show that, if X is a random variable which has the property of last part with parameter σ^2(note that X is not necessarily Gaussian), then for each $t > 0$ we have:

$$\mathbb{P}[|X| \geq t] \leq 2\exp(-\frac{t^2}{2\sigma^2}).$$

3.4 Expectable

Show that for any random variable X, we have:

$$\mathbb{E}[\max(0, Z)] = \int_0^\infty \mathbb{P}[Z \geq x]dx.$$

Using the above formula(and definitely a lot of other tools!) a very useful fact about Normal(and more generally, Gaussian) random variables can be proved. Fortunately you are not going to prove it here! But we state it here since it can give you a very good intuition about this kind of random variables for the rest of the course.

Fact: for a set of n Normal random variables, their expected maximum, is of order $\sqrt{\log n}$. Or in other words, we have the following:

$$c\sqrt{\log n} \leq \mathbb{E}[\max_{1 \leq i \leq n} X_i] \leq C\sqrt{\log n}$$

3.5 $*$Multivariate Gaussian

Suppose that y is a *Gaussian vector*, in other words we have:

$$y \sim \mathcal{N}(\mu, \Sigma)$$

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Show the following statements:

$$p(y_2) = \mathcal{N}(\mu_2, \Sigma_{22})$$

$$p(y_1|y_2) = \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

3.6 Conditional multivariate Gaussian

Take

$$\mathbf{z} \in \mathcal{R}^D, \mathbf{y} \in \mathcal{R}^K, \mathbf{W} \in \mathcal{R}^{K \times D}, \mathbf{b} \in \mathcal{R}^K$$

if we have

$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{W}z + \mathbf{b}, \sigma_\mathbf{y})$$

Show the following statements:

$$p(\mathbf{z}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_\mathbf{z} \\ \mathbf{W}\boldsymbol{\mu}_\mathbf{y} + \mathbf{b} \end{pmatrix}$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_\mathbf{z} & \boldsymbol{\Sigma}_\mathbf{z}\mathbf{W}^T \\ \mathbf{W}\boldsymbol{\Sigma}_\mathbf{z} & \boldsymbol{\Sigma}_y + \mathbf{W}\boldsymbol{\Sigma}_\mathbf{z}\mathbf{W}^T \end{pmatrix}$$
$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}}, \Sigma_{\mathbf{z}|\mathbf{y}})$$
$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}}[\mathbf{W}^T\boldsymbol{\Sigma}_\mathbf{y}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\mathbf{z}]$$
$$\boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}}^{-1} = \Sigma_\mathbf{z}^{-1} + \mathbf{W}^T\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{W}$$

3.7 Gaussian Mixture models

Now assume that in the previous part, the prior distribution is a mixture of K gaussian distributions (GMM), that is, $p(z) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, for which we clearly have $\sum_{k=1}^K \pi_k = 1$. Prove the posterior distribution is another GMM, and calculate its parameters. Hint to avoid a common mistake: The posterior coefficients π'_k are not equal to π_k.

4 Estimators are the Key!

Suppose we have a random vector $X \in \mathbb{R}^d$. All elements are assumed to be *i.i.d* random variables. Assume that we have an observation x. We want to fit a probability distribution to this data and we are going to use the *Maximum Likelihood Estimator* for that.

4.1 MLE 1

Assume that each X_i is a Bernoulli random variable, i.e., $p_{X_i} = \theta^{x_i}(1-\theta)^{1-x_i}$. Also assume that we have observed m ones and k zeros. Find the distribution parameter θ.

4.2 MLE 2

Assume that each X_i is a Normal random variable, i.e., $p_{X_i} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$. Find the mean and variance of the distribution.

4.3 Bias-Variance

Show that for any estimator $\hat{\theta}$ of the parameter θ, we have the following:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}] + (\mathbb{E}[\hat{\Theta}] - \theta)^2$$

4.4 Linear Regression

Consider the following *Linear Regression* model.

$$Y_i = ax_i + b + Z_i$$

Z_i's are $i.i.d$ and of $\mathcal{N}(0, \sigma^2)$ distribution. We know the value of σ and we are given n data like $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$. Using MLE, say how can we estimate $\hat{a}, \hat{b}$.(No calculations is needed for this part)

4.5 Blind estimation

We are given $X_1, X_2, \ldots, X_n$ independent samples from X distribution with mean μ and $Var[X] = \sigma^2$. We want to do an ϵ-accurate estimation of μ. Which means that we want our estimation to be in the $(\mu - \epsilon, \mu + \epsilon)$ range. Show that for an ϵ-accurate estimation, if we have $n = \mathcal{O}(\frac{\sigma^2}{\epsilon^2})$, then with probability at least $\frac{3}{4}$ we will reach our goal.

5 Eigenvalues

Assume $\mathbf{A}$ is a 2×2 matrix with λ_1 and λ_2 being its eigenvalues. If $\lambda_1 \neq \lambda_2$, prove:

$$e^{\mathbf{A}} = \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I + \frac{e_1^{\lambda} - e_2^{\lambda}}{\lambda_1 - \lambda_2} \mathbf{A}$$

6 SVD Decomposition

If the SVD decomposition of matrix $\mathbf{A}$ is defined as $\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^T$, then the pseudo-inverse matrix is defined as $\mathbf{A}^\dagger = \mathbf{V}\Lambda^{-1}\mathbf{U}^T$.

6.1

Show that if A has full row rank, then we have: $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ and if it has full column rank, then we have: $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$. The invertibility of square matrices does not need to be proven.

6.2

Find the SVD decomposition of the matrix $\mathbf{A} = \begin{bmatrix} 1 & 3 & 1 \\ 2 & -1 & 2 \end{bmatrix}$.

7 Vector differentiation

Prove the following vector differentiation formulas.

7.1

$$\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{a}) = \mathbf{a}$$

7.2

$$\nabla_{\mathbf{x}}(\mathrm{Tr}\{\mathbf{x}\mathbf{x}^T\mathbf{A}\}) = \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

7.3 Gradient without explicit differentiation!

Another method to calculate gradients without explicit differentiation is using an equivalent definition of gradients:
For small $\mathbf{\Delta X}$, we have:

$$f(\mathbf{X}+\mathbf{\Delta X}) - f(\mathbf{X}) \approx < \nabla f, \mathbf{\Delta X} > \quad (1)$$

Using this equation and the fact that in the matrix space, the Frobenius inner product is defined as $< \mathbf{A}, \mathbf{B} >= \mathrm{Tr}\left\{\mathbf{A}^T\mathbf{B}\right\}$,prove the following for a symmetric matrix $\mathbf{X}$:

$$\nabla_{\mathbf{X}} - \log\det\{\mathbf{X}\} = -\mathbf{X}^{-2}$$

Hint: You may need an eigenvalue decomposition somewhere in your solution.

7.4 Gradient without explicit differentiation! Part 2

Using the method of the previous part, prove the following:

$$\nabla_{\mathbf{X}}\mathrm{Tr}\left\{\mathbf{X}^{-1}\mathbf{A}\right\} = -\mathbf{X}^{-T}\mathbf{A}^T\mathbf{X}^{-T}$$

Hint: An asymmetric matrix is not always diagonalizable! Use another method for the difference of matrices in your solution.

8 Matrix Frobenius Norm

We define the Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, as follows.

$$\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}\left\{\mathbf{A}^T\mathbf{A}\right\}}$$

8.1

Prove $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2}$

8.2

Given singular values of matrix $\mathbf{A}$: $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{r}\sigma_i(\mathbf{A})^2}$

8.3

Conclude $\max\{\sigma_i(\mathbf{A})\} \leq \|\mathbf{A}\|_F \leq \sqrt{r}\max\{\sigma_i(\mathbf{A})\}$

9 Right or wrong!

Determine the correctness or incorrectness of the following items with sufficient reasoning.

9.1

If $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ are matrices with full rank and $\mathbf{AB} = 0$, then we have: $p + m \leq n$.

9.2

If for some integer values $k \geq 1$ we have $\mathbf{A}^k = 0$, then $\mathbf{A} - \mathbf{I}$ is a matrix with full rank.

9.3

For every $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, the eigenvectors of $\mathbf{AB}$ are equal to the eigenvectors of $\mathbf{BA}$.

10 Calculating normalized eigenvectors from eigenvalues!

We wish to prove the following identity, for diagonalizable matrix $\mathbf{A}$, with diagonalization $\mathbf{A} = \sum_{i=1}^{n} \lambda_i(\mathbf{A}) v_i(\mathbf{A}) v_i(\mathbf{A})^H$, where $\lambda_i(\mathbf{A})$, $1 \leq i \leq n$ are the eigenvalues and $v_i(\mathbf{A})$ are, $1 \leq i \leq n$ are the corresponding eigenvectors, where all eigenvalues are non-zero and the matrix has a simple spectrum.

$$|v_{i,j}(\mathbf{A})|^2 \prod_{\substack{k=1 \\ k \neq i}}^{n} (\lambda_i(\mathbf{A}) - \lambda_k(\mathbf{A})) = \prod_{k=1}^{n-1} (\lambda_i(\mathbf{A}) - \lambda_k(\mathbf{M_{jj}}))$$

Where $\mathbf{M_{ij}}$ is the submatrix formed by removing the ith row and jth column from the original matrix, $\mathbf{A}$.
To do this we take the following steps.

10.1

For any square matrix S, with eigenvalues $\lambda_1(\mathbf{S}), \lambda_2(\mathbf{S}), \ldots, \lambda_n(\mathbf{S})$, prove $\det(\mathbf{S}) = \prod_{i=1}^{n} \lambda_i(\mathbf{S})$

10.2

Prove $\mathbf{A}\text{adj}(\mathbf{A}) = \det(\mathbf{A})I = \text{adj}(\mathbf{A})\mathbf{A}$, in which $\text{adj}(\mathbf{A})$ is the adjugate matrix of $\mathbf{A}$. We define the coefficients of $\text{adj}(\mathbf{A})$ by $(\text{adj}(\mathbf{A}))_{ij} = (-1)^{i+j}(\mathbf{M_{ji}})$

10.3

Show that $\text{adj}(\mathbf{A})$ has the diagonalization $\text{adj}(\mathbf{A}) = \sum_{i=1}^{n} \left(\prod_{k=1;k\neq i}^{n} \lambda_k(\mathbf{A}) \right) v_i(\mathbf{A}) v_i(\mathbf{A})^H$

10.4

Now prove the identity!

11 Optimization

In the following lessons, you will become familiar with various classifiers, one of which is Support Vector Machine or SVM for short. In this question, we aim to examine this classifier for inseparable data. As you will see later, to find the best classifier, we will encounter an optimization problem with inequality constraints as follows.

$$\begin{cases} \min\limits_{\mathbf{w},w_0,\eta} \quad J(\mathbf{w}, w_0, \eta) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{N} \eta_i \\ \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x_i} + w_0) \geq 1 - \eta_i \quad , \quad \eta_i \geq 0 \qquad i = 1, 2, \ldots, N \end{cases}$$

11.1

Formulate the Lagrangian for the above problem.

11.2

Obtain the solution to the problem by applying the Karush-Kuhn-Tucker (KKT) conditions.