

Introduction to Machine Learning (67577)

Exercise 1 Mathematical Background

Second Semester, 2021

Contents

1	Theoretical	2
1.1	Linear Algebra	2
1.1.1	Recap	2
1.1.2	Singular Value Decomposition	2
1.2	Multivariate Calculus	2
2	Practical	3
2.1	Multivariate Gaussians	3
2.2	Concentration Inequalities	3

1 Theoretical

1.1 Linear Algebra

1.1.1 Recap

1. Calculate the projection of $v = (1, 2, 3, 4)^\top$ on the vector $w = (0, -1, 1, 2)^\top$.
2. Calculate the projection of $v = (1, 2, 3, 4)^\top$ on the vector $w = (1, 0, 1, -1)^\top$.
3. Prove the angle between two non-zero vectors $v, w \in \mathbb{R}^m$ is $\pm 90^\circ$ iff $v^\top w = 0$.
4. Prove that orthogonal matrices are isometric transformations. That is let $T : V \mapsto W$ be some linear transformation and A the corresponding matrix. Then if A is orthogonal then $\forall x \in V \ ||Ax|| = ||x||$.

1.1.2 Singular Value Decomposition

5. Let A be an invertible matrix. Write a formula for the inverse of A using only the matrices of the SVD of A : $U\Sigma V^\top$. Explain why knowing the SVD of matrix is useful in this context.
6. Compute the SVD of

$$A = U\Sigma V^\top = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

That is, find the matrices U, Σ, V^\top where U, V are orthogonal matrices and Σ is diagonal. Guidance:

- Calculate $A^\top A$
 - Deduce V and Σ . Use the *Eigenvalues Decomposition*, either manually (good practice) or use `numpy.linalg.eig`
 - Find U using the equality $AV = U\Sigma$
7. (Power Iteration) In this section we will implement an algorithm for SVD decomposition, we will use the relation between SVD of A to EVD of $A^\top A$ that we saw in recitation. For some $A \in M_{m \times n}(\mathbb{R})$, define $C_0 = A^\top A$. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues of C_0 , with the corresponding normalized eigenvectors v_1, v_2, \dots, v_n , ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Assume $\lambda_1 > \lambda_2$, where λ_1 is the largest eigenvalue and λ_2 is the second-largest one. Define $b_{k+1} = \frac{C_0 b_k}{\|C_0 b_k\|}$, and let $b_0 = \sum_{i=1}^n a_i v_i$, where $a_1 \neq 0$. Show that: $\lim_{k \rightarrow \infty} b_k = \pm v_1$.

1.2 Multivariate Calculus

8. Let $x \in \mathbb{R}^n$ be a fixed vector and $U \in \mathbb{R}^{n \times n}$ a fixed orthogonal matrix. Calculate the Jacobian of the function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$f(\sigma) = U \cdot \text{diag}(\sigma) U^\top x$$

Where $\text{diag}(\sigma)$ is an $n \times n$ matrix where

$$\text{diag}(\sigma)_{ij} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

9. Use the chain rule to calculate the gradient of $h(\sigma) = \frac{1}{2} \|f(\sigma) - y\|^2$
10. Calculate the Jacobian of the softmax function (initial steps can be found in recitation file):

$$g(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

2 Practical

2.1 Multivariate Gaussians

In this question we will examine the three-dimensional case. We want to show how linear transformations affect the data set and in result the covariance matrix. First we will generate random points with mean values at the origin and unit variance $\sigma(x) = \sigma(y) = \sigma(z) = 1$.

11. Download the file `3d_gaussian.py`. Use the identity matrix as the covariance matrix to generate random points and then plot them (with the given function). *You are not obligated to use or submit the “3d_gaussian.py” file. It is there to help you start*
12. Transform the data with the following scaling matrix:

$$S = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Plot the new points. What does the covariance matrix look like now (both analytically and numerically)?

13. Multiply the scaled data by random orthogonal matrix. Plot the new points. What does the covariance matrix look like now?
14. In recitation we claimed that the marginal distribution of a gaussian is still gaussian. Plot the projection of the data to the x, y axes. What does it look like? Add the plot to the submission.
15. In recitation we claimed that the conditional distribution of a gaussian is still gaussian. Only for points where $0.1 > z > -0.4$: Plot the projection of the points to the x, y axes. What does it look like? Add the plot to the submission.

2.2 Concentration Inequalities

In the following weeks we will introduce the PAC model for learnability, that tells us when a given hypothesis class is learnable. It describes a relationship between 3 factors: the sample complexity (m , number of samples in our dataset), our approximation of the learner's loss (ϵ) and our confidence level (δ). **Intuitively speaking**, given a dataset of size m drawn from some distribution \mathcal{D} , we want "with high probability" $(1 - \delta)$ that the learner's loss over that dataset will be very small (ϵ). In the following question we will show some aspects of the relationship between the factors. You will not need (yet) any deep understanding of this model, but just the probabilistic bounds seen in class.

16. Consider the following scenario. Let X be a Bernoulli random variable representing a coin toss, with probability of ϵ being 1. You are given 100,000 sequences of 1,000 coin tosses (arranged in a matrix, “data”, of 100000 rows and 1000 columns. That is, each coin toss is like:

$$X \sim \text{Ber}(\epsilon)$$

and all coin tosses can be thought of as:

$$\forall j \in [100000] \quad X_1^j, \dots, X_{1000}^j \stackrel{iid}{\sim} \text{Ber}(\epsilon)$$

where *iid* means that all have the same distribution (Bernoulli ϵ) and are independent (X_l^j, X_k^j are independent $\forall l \neq k$). In addition tosses of different sequences (and therefore the sequences too) are independent.

Using this scenario we will explore the Chebyshev and Hoeffding bounds on our estimation of ε . Define a variable `epsilon` which gets the values $[0.5, 0.25, 0.1, 0.01, 0.001]$. For each value of `epsilon` generate a dataset and perform the tasks below. To generate the data use the commands:

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

- (a) For the first 5 sequences of 1000 tosses (the first 5 rows in “data”), plot the estimate $\bar{X}_m = \frac{1}{m} \sum_{i=1}^m X_i$ 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 seen in class) and for each ε , plot the upper bound on $\mathbb{P}_{X_1, \dots, X_m} (|\bar{X}_m - \mathbb{E}[X]| \geq \varepsilon)$ (derived in class) as a function of m (where m ranges from 1 to 1000). X refers to the random variable seen in the explanation above.

You are expected to create 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 $|\bar{X}_m - \mathbb{E}[X]| \geq \varepsilon$ as a function of m , where this time $\mathbb{E}[X] = p = 0.25$. What are you expecting to see in these plots? Explain. In your submission be sure to add plots of clause (b) and of clause (c).

¹if the bound calculated is greater than 1, you should substitute it with 1. This is because we are bounding a probability which is always smaller than 1.