COMS 4771 Practice Problems

Last updated: 2019-10-04 17:36:38-04:00

These practice problems are for practice only; they are not to be submitted. Please feel free to discuss on Piazza, in office hours, etc.

We will not be posting solutions.

Contents

1 Homework 0 review 2
   1.1 Question 6 - Symmetric matrix 2
   1.2 Sum of independent Gaussian 2

2 Probability 2
   2.1 Markov 2
   2.2 MLE 3
   2.3 MLE 3
   2.4 Bayes Optimal Classification 3

3 Decision trees 3
   3.1 Binary trees 3
   3.2 Build your tree 3

4 Linear algebra 4
   4.1 Linear Algebra in Linear Regression 4
   4.2 Distance optimization 4

5 Nearest Neighbor 4
   5.1 Convexity of nearest neighbor regions 4

6 Linear regression 5
   6.1 Heteroskedastic noise in linear regression 5
      6.1.1 Part (a) 5
      6.1.2 Part (b) 5
   6.2 Linear algebraic perspective 5
      6.2.1 Part (a) 5
      6.2.2 Part (b) 5
      6.2.3 Part (c) 5
   6.3 The least (Euclidean) norm solution 5

7 Linear algebra practice 6
   7.1 Gradients and derivatives 6
1 Homework 0 review

1.1 Question 6 - Symmetric matrix

Let \( g: \mathbb{R}^2 \to \mathbb{R} \) be the function defined by \( g(x) := \frac{1}{2} x^\top A x - b^\top x + c \) where

\[
A := \begin{bmatrix} 3 & 5 \\ 5 & 3 \end{bmatrix}, \quad b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \text{and} \quad c := 9.
\]

1. Compute the determinant of \( A \)

2. Without any calculations, show that the eigenvalues of \( A \) are real and of opposite sign.

3. Show that \( b^\top x = O_{x \to \infty}(||x||) \). See footnotes for \( O \) materials.

4. Show that \( g \) has neither a minimum nor a maximum

1.2 Sum of independent Gaussian

Let \( X \sim \mathcal{N}(0, 1) \) and \( Y \sim \mathcal{N}(0, 1) \) be two independent normal random variables

1. Show that \( X + Y \) is also a normal random variable

2 Probability

2.1 Markov

Show that:

1. For any (measurable) event \( A \), we have \( \Pr[A] = \mathbb{E}[1_{\{A\}}] \)

2. For any non-negative random variable \( X \), and any \( c > 0 \),

\[
\Pr[X \geq c] \leq \frac{\mathbb{E}[X]}{c}
\]

(Hint: compare the output of the function \( 1_{\{X > c\}} \) with the outcome of \( X \).)

3. Let \( X_1, \ldots, X_n \in \{0, 1\} \) be the outcome of \( n \) coin flips drawn from \( \text{Bern}(p) \). We can estimate the value of \( p \) using the sample mean \( \sum X_i/n \). Use the previous result to show that

\[
\Pr \left[ \left| \sum X_i/n - p \right| > \epsilon \right] \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}
\]

(Hint: \( X > 0 \implies \Pr[X < c] = \Pr[X^2 < c^2] \))

2.2 MLE

Let \( X_1, \ldots, X_n \in \{1, 6\} \) be the outcomes of \( n \) i.i.d. rolls of a (potentially) weighted die, i.e., from \( \text{Categorical}(p_1, p_2, p_3, p_4, p_5, p_6) \). The probability of the observed rolls is

\[
\prod_{i=1}^{n} \sum_{j=1}^{6} 1[X_i = j] p_j
\]

where \( \sum p_i = 1 \). Show that the MLE estimator for \( p_j \) is the expected \( \#X_j/n \), i.e. the fraction of rolls with that face value. Hint: compute the log probability and make sure to enforce the \( \sum p_i = 1 \) condition (for example, with Lagrange multipliers) while maximizing.

---

1 If you want precision on \( O \) notations, you can check the Wikipedia page [https://en.wikipedia.org/wiki/Big_O_notation](https://en.wikipedia.org/wiki/Big_O_notation)

2 Paragraph 0.3 of [https://people.eecs.berkeley.edu/~vazirani/algorithms/chap0.pdf](https://people.eecs.berkeley.edu/~vazirani/algorithms/chap0.pdf) also explain the link with algorithms complexity analysis.
2.3 MLE
Consider the probability density \( p(x) = 2\theta xe^{-x^2\theta} \) for \( x \geq 0 \) and \( p(x) = 0 \) for \( x < 0 \). Find the maximum likelihood estimator for \( \theta \) given an iid sample \((x_1, \ldots, x_n)\).

2.4 Bayes Optimal Classification
In classification, the loss function we usually want to minimize is the 0/1 loss:

\[ l(f(x), y) = \mathbb{1}\{f(x) \neq y\} \]

where \( f(x), y \in \{0, 1\} \) (i.e., binary classification). In this problem we will consider the effect of using an asymmetric loss function:

\[ l_{\alpha,\beta}(f(x), y) = \alpha \mathbb{1}\{f(x)=1, y=0\} + \beta \mathbb{1}\{f(x)=0, y=1\}. \]

Under this loss function, the two types of errors receive different weights, determined by \( \alpha, \beta > 0 \). Determine the Bayes optimal classifier, i.e., the classifier that achieves minimum risk assuming the distribution of the random example \((X,Y)\) is known, for the loss \( l_{\alpha,\beta} \) where \( \alpha, \beta > 0 \).

3 Decision trees
3.1 Binary trees
Suppose you have a decision tree where the splitting rule can have more than two possible output values, so that non-leaf nodes can have more than two children. Show how to convert it into a decision tree in which the splitting rules have only two possible output values.

3.2 Build your tree
We recall two measures of uncertainty \( M \in \{G, E\} \) for decision trees at node \( n \) where \( p_k \) is the proportion of examples reaching node \( n \) with label \( k \):

\[ \text{Gini index at node } n: \quad G(n) = 1 - \sum_k p_k^2 \]  
\[ \text{Entropy at node } n: \quad E(n) = -\sum_k p_k \log p_k. \]  

(1) \hspace{1cm} (2)

For one node \( n \) with children \( l_{\text{left}}, l_{\text{right}} \), we recall the overall uncertainty:

\[ \sum_{l \in \{l_{\text{left}}, l_{\text{right}}\}} M(l) \cdot (# \text{training examples reaching } l) \]

1. Show that the Gini index rewrites: \( \sum_k p_k(1 - p_k) \)
2. Compute both measures for:
   (a) All the observations of figure [-figure]
   (b) The observations only with \( x > -1 \)
   (c) The observations only with \( x < 1 \)
3. Using only splits of the form \((x_1, x_2) \rightarrow \mathbb{1}\{x_i > c\} \) for \( c \in \mathbb{R} \), build 2 binary trees of height 1 that minimize the overall uncertainty (one using the Gini index, and one using the entropy).
4. Do you think using 2 other forms of splits instead, you can build a tree with less nodes?
4 Linear algebra

4.1 Linear Algebra in Linear Regression

Consider the a linear model $y = X\beta + \epsilon$, and the fitted values from linear regression $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty$. For simplicity, let’s assume $\epsilon_i$’s are i.i.d. normal with 0 mean and $\sigma^2$ variance. Recall an alternative definition of the degrees of freedom:

$$df(\hat{f}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(y_i, \hat{y}_i)$$

What is the theoretical degrees of freedom for this linear regression?

Solution

Denote $H = X(X^TX)^{-1}X^T$, then $\hat{y} = Hy$. Now, $\text{cov}(\hat{y}, y) = \text{cov}(Hy, y) = H\text{cov}(y, y) = H(\sigma^2 I_{n\times n}) = \sigma^2 H$.

Thus, $df(\hat{f}) = \frac{1}{\sigma^2} \text{Trace}(\text{cov}(\hat{y}, y)) = \text{Trace}(H)$ where $H = X(X^TX)^{-1}X^T$.

4.2 Distance optimization

Given a point $p \in \mathbb{R}^d$ and a plane $D := \{x \in \mathbb{R}^d | w \cdot x = 0\}$ defined as the set of vectors orthogonal to a vector $w$, derive a formula for the minimum distance between $p$ and any point on $D$.

5 Nearest Neighbor

5.1 Convexity of nearest neighbor regions

1. Given a collection of labeled examples $D := \{(x_i, y_i)\}_{i=1}^{n}$, and two unlabeled examples $t_1, t_2$, suppose that the $t_1$ and $t_2$ have the same nearest neighbor in $D$, $d_1$ (when using $\ell_2$ norm). Prove that for any point lying on the line segment in between $t_1$ and $t_2$, that point’s nearest neighbor in $D$ will also be $d_1$.

2. If $t_1$ and $t_2$’s nearest neighbors in $D$ simply have the same training label (but may be different examples), must the nearest neighbor of every point on the line segment in between $t_1$ and $t_2$ have that training label as well?
6 Linear regression

6.1 Heteroskedastic noise in linear regression

6.1.1 Part (a)
Let $P_{\beta}$ be a probability distribution on $\mathbb{R}^d \times \mathbb{R}$ for the random pair $(X,Y)$ (where $X = (X_1, \ldots, X_d)$) such that $X_1, \ldots, X_d \sim i.i.d.$ $N(0,1)$, and $Y | X = x \sim N(x^T \beta, \|x\|_2^2)$, $x \in \mathbb{R}^d$.

Here, $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ are the parameters of $P_{\beta}$.

True or false: The linear function with the smallest squared loss risk with respect to $P_{\beta}$ is $\beta$.

Answer with "true" or "false", and briefly (but precisely) justify your answer.

6.1.2 Part (b)
Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be given, and assume $x_i \neq 0$ for all $i = 1, \ldots, n$. Let $f_{\beta}$ be the probability density function for $P_{\beta}$ as defined in Part (a). Define the function $Q : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$Q(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln f_{\beta}(x_i, y_i), \quad \beta \in \mathbb{R}^d.$$ 

Find a system of linear equations $A\beta = b$ over variables $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ such that its solutions are maximizers of $Q$ over all vectors in $\mathbb{R}^d$.

Write the system of linear of equations by defining the left-hand side matrix $A \in \mathbb{R}^m \times \mathbb{R}^d$ and right-hand side vector $b \in \mathbb{R}^m$ (here, $m$ is the number of linear equations), and briefly (but precisely) justify your answer. You may define $A$ and $b$ as products of matrices and vectors if you like, but make sure these matrices and vectors are also clearly defined.

6.2 Linear algebraic perspective

Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be given, and let $\hat{\mathcal{R}}$ be defined by $\hat{\mathcal{R}}(\beta) := \|A\beta - b\|_2^2$.

6.2.1 Part (a)
Suppose the rank of $A$ is smaller than $d$. Is the minimizer of $\hat{\mathcal{R}}$ to be uniquely defined?

Answer with "yes" or "no", and briefly (but precisely) justify your answer.

6.2.2 Part (b)
Suppose the rank of $A$ is smaller than $d$. Is the orthogonal projection of $b$ onto the range of $A$ uniquely defined?

Answer with "yes" or "no", and briefly (but precisely) justify your answer.

6.2.3 Part (c)
Suppose you only have $A$ and not $b$, but you are given an orthogonal projection $\hat{b}$ of $b$ onto the range of $A$. Explain how to find a minimizer of $\hat{\mathcal{R}}$ using only $A$ and $\hat{b}$.

Answer with precise pseudocode for a procedure that takes as input $A$ and $\hat{b}$ and returns a minimizer of $\hat{\mathcal{R}}$. Briefly (but precisely) justify the correctness of the procedure.

6.3 The least (Euclidean) norm solution

Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. In lecture, we saw that the solution to the normal equations of minimum Euclidean norm is an element of the row space of $A$. Prove that if $w$ is a solution to the normal equations that also is in the row space of $A$, then it is unique.
7 Linear algebra practice

7.1 Gradients and derivatives

In this exercise, \( A \in \mathbb{R}^{n \times n} \), \( b \in \mathbb{R}^n \), \( J \) is the Jacobian and \( \nabla \) is the gradient. Gradient is defined for scalar function from \( \mathbb{R}^n \) to \( \mathbb{R} \) and Jacobian for vector-valued functions. Sometime people tend to say gradient even if the function is vector-valued.

(a) Let \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \), compute \( J(f)(x) \) for any \( x \in \mathbb{R}^n \) where \( f \) is differentiable.

(b) Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), compute \( \nabla f(x) \) for any \( x \in \mathbb{R}^n \) where \( f \) is differentiable.

(c) Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), compute \( \nabla f(x) \) for any \( x \in \mathbb{R}^n \) where \( f \) is differentiable.

(d) Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), compute \( \nabla f(x) \) for any \( x \in \mathbb{R}^n \) where \( f \) is differentiable.