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

Contents

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

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

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

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

5 Nearest Neighbor
   5.1 Convexity of nearest neighbor regions .................................................. 4
1 Homework 0 review

1.1 Question 6 - Symmetric matrix

Let \( g : \mathbb{R}^2 \to \mathbb{R} \) be the function defined by \( g(x) := \frac{1}{2} x^\top Ax - b^\top x + c \) 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. \[^1\]
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 \( \Pr[A] = \mathbb{E}[\mathbb{1}_{\{A\}}] \)
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 \( \mathbb{1}_{\{X > c\}} \) 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 \url{https://en.wikipedia.org/wiki/Big_O_notation}
\[^2\]Paragraph 0.3 of \url{https://people.eecs.berkeley.edu/~vazirani/algorithms/chap0.pdf} also explain the link with algorithms complexity analysis.
2.3 MLE

Consider the probability density \( p(x) = 2\theta xe^{x^2} \) 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:

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

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:

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

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 \( k \):

\[
 Gini index at node n: \ G(n) = 1 - \sum_k p_k^2 \tag{1}
\]

\[
 Entropy at node n: \ E(n) = -\sum_k p_k \log p_k. \tag{2}
\]

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 1
   (b) The observations only with \( x > -1 \)
   (c) The observations only with \( x < 1 \)

3. Using only splits of the form \((x_1, x_2) \mapsto 1\{x_i > c\} \) 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 linear model $y = X\beta + \epsilon$, and the fitted values from linear regression $\hat{y} = X\hat{\beta} = X(X^T X)^{-1}X^T y$. 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 $H = X(X^T X)^{-1}X^T$, 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 $H = X(X^T X)^{-1}X^T$.

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?