CSCC11 Winter 2025 Assignment 2 TP1

Parameter Estimation, KNN, and Decision Trees

Due: March 4th, 2025, 22:00 EDT

This is the first theory part of assignment 2, making use of lectures and tutorials for the first six weeks. You are asked to derive solutions to problems to demonstrate your understanding of the concepts learned. Submission instructions can be found at the end of the handout. We remind you that the work you hand in must be your own.

Please see the Generative AI Policy section regarding using AI chatbots such (ChatGPT, Copilot, Gemini et. al.) to help complete assignments in this course.

Student Name: Your Name Here

Student ID: Your student ID number

Email: Your UofT Email

Theory Part

Q1: MAP and Bayes Estimates (11/70 points)

In the lecture, we discussed parameter estimation for a Bernoulli random variable from N independent observations. In particular, we estimated the probability θ that a coin toss will land heads-side up, given the outcomes of N coin tosses. Initially, we assumed a uniform prior distribution over the potential value of θ. Now, suppose we have a non-uniform prior, such as a Beta distribution:

$$\text{Beta}(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where

- θ is the variable of interest, typically representing a probability.
- $\alpha, \beta > 0$ are a shape parameters.
- $B(\alpha, \beta)$ is the Beta function, serving as a normalization constant ensuring that the density integrates to 1.

(a) [2 pts] Setting $\alpha = \beta = 1$ in the Beta distribution, what function will we obtain? How about setting $\alpha = \beta = 2$? For both cases, find the maximum value of the probability density function and identify the location of this maximum.

(b) [2 pts] Suppose we observe N independent coin flips. Derive a mathematical expression for the posterior distribution over θ using the Beta prior.

(c) [2 pts] Derive a mathematical expression for the MAP estimate for θ.

(d) [2 pts] Compare two different estimators, namely,

(i) θ_{MAP} with a Beta prior, and

(ii) the Bayes estimate with a uniform prior.

How do they differ? Is it possible to obtain θ from the Bayes estimate using a Beta prior?

(e) [3 pts] Discuss the differences in estimates of θ from

(i) the MAP estimate with a uniform prior,

(ii) the Bayes estimate with the uniform prior, and

(iii) the MAP estimate with a Beta prior,

when the number of observed coin tosses is small, versus when the number of coin tosses tends to infinity. In a couple of sentences, explain which one seems more useful.

Q2. KNN (10/70 points)

In class, we claimed that points are all far apart in high-dimensional space and they tend to have equidistance. This means that the differences in distance between nearby and faraway points become negligible, which can significantly reduce the effectiveness of distance-based algorithms such as KNN. This curse of dimensionality is explained geometrically in the textbook [CB] Section 1.4.

You will prove this curse of dimensionality in this question by analyzing the properties of the squared Euclidean distance from a probabilistic perspective. Let $U, V \in \mathbb{R}^d$ be two random vectors in the high-dimensional space. Assume they are independent, where each coordinate U_j and V_j is drawn independently from the uniform distribution on the internval $[0, 1]$,where $j = 1, 2, \ldots, d$. Define squared Euclidean distance variable

$$D = \|U - V\|^2 = \sum_{j=1}^{d} D_j = \sum_{j=1}^{d} (U_j - V_j)^2$$

(a) [1 pts] Derive the expected squared Euclidean distance $\mathbb{E}[(U_j - V_j)^2]$ for each coordinate between U_j and V_j, where $j = 1, 2, \ldots, d$.

(b) [2 pts] Derive the expected squared Euclidean distance $\mathbb{E}[\|U - V\|^2]$ between U and V. Show that the expected squared Euclidean distance is a linear function of d. Hence when d is large, the points are far apart.

(c) [1 pts] Compute the variance of the squared Euclidean distance $\texttt{Var}[(U_j - V_j)^2]$for each coordinate, where $j = 1, 2, \ldots, d$.

(d) [2 pts] Show that the variance of the expected squared Euclidean distance $\texttt{Var}[\|U - V\|^2]$ is proportional to d. Hint: the variance of the sum of independent random variables is the sum of their variances.

(e) [4 pts] Show that in high dimension, points are centered around the mean, hence almost all pairs have the same distances. Hint: Use Chebyshev's inequality to bound the probability $|D - \mathbb{E}(D)|$ and see how the bound behaves when $d \to \infty$.

$|D - E(D)| \to$ constant when $d \to \infty$

$\downarrow$ bounded by constant $=$

$P(|D - E(D)| \geq k\sqrt{\text{Var}(D)}) \leq \frac{1}{k^2}$

Q3. Decision Trees (9/70 points)

In class, we discussed how to use information gain to determine which attribute (or feature) is the most useful for discriminating between two classes. In this question, we are looking to construct a binary decision tree to classify whether a computer system is considered critical or not based on the following features.

- SSD: Does the system have an SSD installed? (1 = Yes, 0 = No)
- UPS: Does the system have an Uninterruptible Power Supply? (1 = Yes, 0 = No)
- ServerType: The type of server: Application, Web, or Database.

The training dataset in `a2_compsys.csv` (see Quercus).

(a) [1 pts] For the root node of the tree, compute the entropy for the distribution over the target variable (i.e., Diagnosis) for which we aim to learn a prediction tool.

(b) [2 pts] Prove that entropy (for discrete random variables) is always non-negative.

(c) [3 pts] How many different split functions are there for this dataset? What are they? For each, compute the information gain at the root node.

(d) [1 pt] What split function would you use for the root of the decision tree?

(e) [2 pts] Draw (by hand if you like) the full decision tree that you would expect to learn.

Submission Instructions

The theory part should be done with text editor. The LaTex source file is available on Querucs. We highly recommend using LaTex, though you are free to use any other text editors. The point is that your answers should be readable. Your answers should be handed in electronically as a single pdf file (for all 3 questions). The file should be called `a2_tp1_solution.pdf`, and it should be submitted to Markus `A2_TP1`.

Generative AI Policy

Asking general questions about concepts relevant to the assignment questions is permitted. You are not permitted to directly ask chatbots for hints or solutions of assignment questions. You should not feed the contents of the assignment into the chatbot and then copy or paraphrase the solution provided by the chatbot. For the programming part of the assignment,

you must properly attribute any code generated by the chatbot. You need to include all chat transcripts (including the prompts) that have helped you to complete the programming assignments in the file named `GenAILog.pdf`.