Proposal

STA2453

Winter 2025

Proposal

Proposal

- 1(-2) page writeup of the project you have chosen
- Explain the question clearly — what are you trying to achieve?
- Explain the data thoroughly
- Include a table and/or 1-2 figures that clearly show important aspects of the data
- Given the aims of the project, outline some tools you may use. Explain why these tools are useful and adequate for the project.
- **Create a timeline for the different parts of the project, e.g. Gantt chart**

Proposal

- Explaining the question clearly:

  - explaining the question to someone who knows a lot about the project

  - explaining the question to someone who knows the technical details but not about the specific project

  - explaining the question to someone who doesn’t know much about the technical details or the project

Proposal

- Explaining the data clearly:

  - What are the response variables? Or how has the problem been quantified?

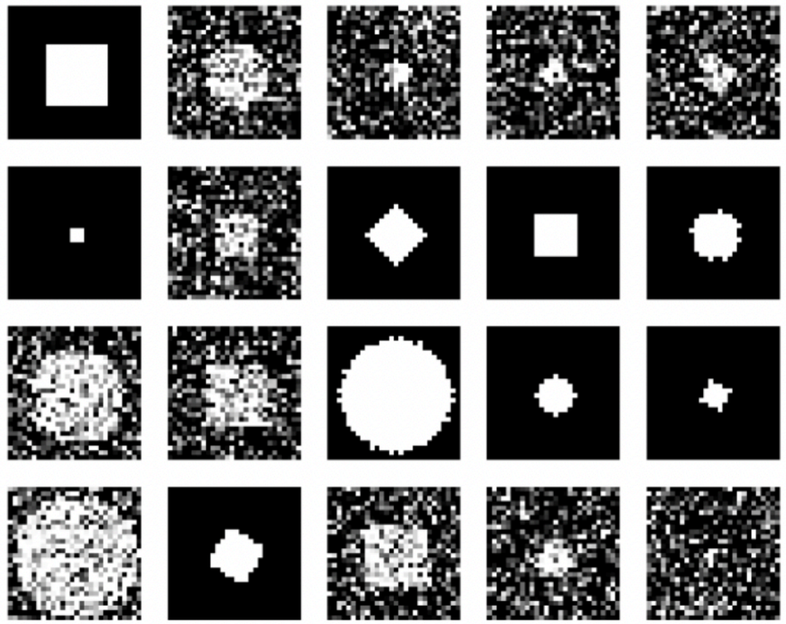
  E.g. if you're working with video data, discuss how the video data has been processed in order to work with numerical values

  - What are some features/covariates you plan to use, what values can they take on?

Proposal

- Include 1-2 tables/figures — these need to connect with questions you have about your data or that connects with the research question of interest

- For example, say you wanted to construct a model to classify circles and squares. You may want to include in the initial proposal what those circles and squares look like.



Identifying approaches to use

- Be ready to answer, “why is the method you used appropriate to answer the question?”
- That involves knowing:

  - the data very well, understanding its structure

  - what the methods do

  - why the data you have is appropriate for the methods of your choosing

Identifying approaches to use

14 Predictive models: an overview

14.1 Introduction

The vast majority of machine learning is concerned with tackling a single problem, namely learning to predict outputs $\boldsymbol{y}$ from inputs $\boldsymbol{x}$ using some function f that is estimated from a labeled training set $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) : n = 1 : N\}$, for $\boldsymbol{x}_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and $\boldsymbol{y}_n \in \mathcal{Y} \subseteq \mathbb{R}^C$. We can model our uncertainty about the correct output for a given input using a conditional probability model of the form $p(\boldsymbol{y}|f(\boldsymbol{x}))$. When $\mathcal{Y}$ is a discrete set of labels, this is called (in the ML literature) a **discriminative model**, since it lets us discriminate (distinguish) between the different possible values of $\boldsymbol{y}$. If the output is real-valued, $\mathcal{Y} = \mathbb{R}$, this is called a **regression model**. (In the statistics literature, the term "regression model" is used in both cases, even if $\mathcal{Y}$ is a discrete set.) We will use the more generic term "**predictive model**" to refer to such models.

Identifying approaches to use

14.2 Evaluating predictive models

In this section we discuss how to evaluate the quality of a trained discriminative model.

14.2.1 Proper scoring rules

It is common to measure performance of a predictive model using a **proper scoring rule** [GR07], which is defined as follows. Let $S(p_{\boldsymbol{\theta}}, (y, \boldsymbol{x}))$ be the score for predictive distribution $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ when given an event $y|\boldsymbol{x} \sim p^*(y|\boldsymbol{x})$, where p^* is the true conditional distribution. (If we want to evaluate a Bayesian model, where we marginalize out $\boldsymbol{\theta}$ rather than condition on it, we just replace $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ with $p(y|\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$.) The expected score is defined by

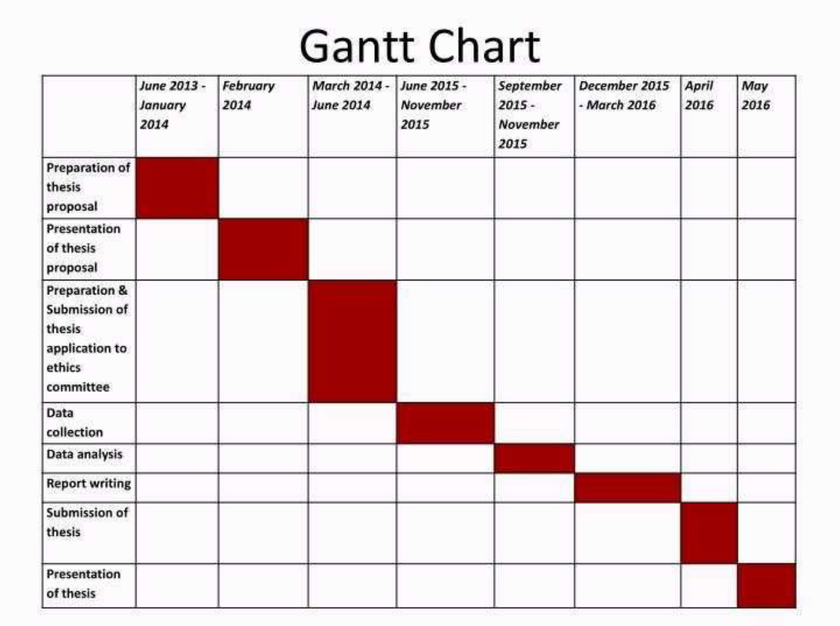
$$S(p_{\boldsymbol{\theta}}, p^*) = \int p^*(\boldsymbol{x}) p^*(y|\boldsymbol{x}) S(p_{\boldsymbol{\theta}}, (y, \boldsymbol{x})) dy d\boldsymbol{x} \tag{14.14}$$

A proper scoring rule is one where $S(p_{\boldsymbol{\theta}}, p^*) \leq S(p^*, p^*)$, with equality iff $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = p^*(y|\boldsymbol{x})$. Thus maximizing such a proper scoring rule will force the model to match the true probabilities.

Proposal

- Create a timeline for what tasks you will accomplish — e.g. via a Gantt chart, kanban board

  Here's an older example:



- Here's something that's more updated using notion: https://notion-templates.notion.site/Team-Projects-13c17954d9c280c19ef2c69951bd98e0

- These are used for project management

Exploratory Data Analysis

https://r4ds.hadley.nz/eda

10 Exploratory data analysis

10.1 Introduction

This chapter will show you how to use visualization and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle. You:

1. Generate questions about your data.
2. Search for answers by visualizing, transforming, and modelling your data.
3. Use what you learn to refine your questions and/or generate new questions.

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive insights that you'll eventually write up and communicate to others.

EDA is an important part of any data analysis, even if the primary research questions are handed to you on a platter, because you always need to investigate the quality of your data. Data cleaning is just one application of EDA: you ask questions about whether your data meets your expectations or not. To do data cleaning, you'll need to deploy all the tools of EDA: visualization, transformation, and modelling.

Final project presentation

Format

- Presentations will be 5 minutes long and pre-recorded
- You will each have 4-5 presentations to watch and be prepared with questions for on the final day
- Last day of class: each person will create a single slide and take questions from the people assigned to them