

School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2021)
Workshop Week 2

Considering the following problems:

- (i) Building a system that guesses what the weather (temperature, precipitation, etc.) will be like tomorrow
- (ii) Predicting products that a customer would be interested in buying, based on other purchases that customer has previously made
- (iii) Skin cancer screening test
- (iv) Automatically identifying the author of a given piece of literature (NLP)
- (v) Finding the best burrito in the United States of America

classification (sunny, rainy etc.)

Regression

hard to generalize
symptom is specific

clustering

image \Rightarrow paragraph describing the pic

biased aged people \checkmark
young people (new)

e.g. recommendation system

1. Identify the "concept" we might attempt to "learn" for each problem (Task Identification)

2. For each problem-task, identify what the "instances" and "attributes" might consist of (Choosing the Data Representative)

(v) purchase history / review / ranking / ingredients /

3. For each problem-task, conjecture whether a typical strategy is likely to use "supervised" or "unsupervised" Machine Learning (Picking a Suitable Model)

pic of burrito itself / packaging etc.

4. For each problem-task, consider how easy or difficult it would be to make a model that generalizes to new cases. For example, could you predict the weather in any city in the world, or just in one specific city?

5. What kinds of assumptions might a machine learning model make when tackling these problems?

For all questions above: majority of training data is corrected / unbiased

(Supervised)

(iv) ① Normal identification:

correlation between attributes & label
the model is capable of processing the new instance
attributes are independent with each other.

② Plagiarism detection: semi-supervised

(Database: online articles \Rightarrow check overlaps)

irrelevant attributes are fine \Rightarrow more parameters.
the model could ignore useless ones so the
training process is faster / more accurate.