

MA 506 Probability and Statistical Inference

Lecture 1: Introduction

1: Definitions

\$Probability\$: Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility of the event and 1 indicates certainty [1].

\$Statistical Inference\$: Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability [2].

2: Why study this topic ?

Probability theory and statistical inference theory are important tools for data scientists and analysts that enables the computation of how certain we are about a numerical prediction. This estimation of reliability informs our level of trust in the output of a computer algorithm

2.1 Which prediction to trust more ?

Running speed was recorded for several days for 5 people. Based on the speed data, assume the computer algorithm came up with the following predictions and uncertainty in estimation of speed for these people. From Jim, Slim, Harry, Dick and Tom, whose prediction would you trust most and whose the least ?

```
In [1]: import matplotlib.pyplot as plt
import numpy as np

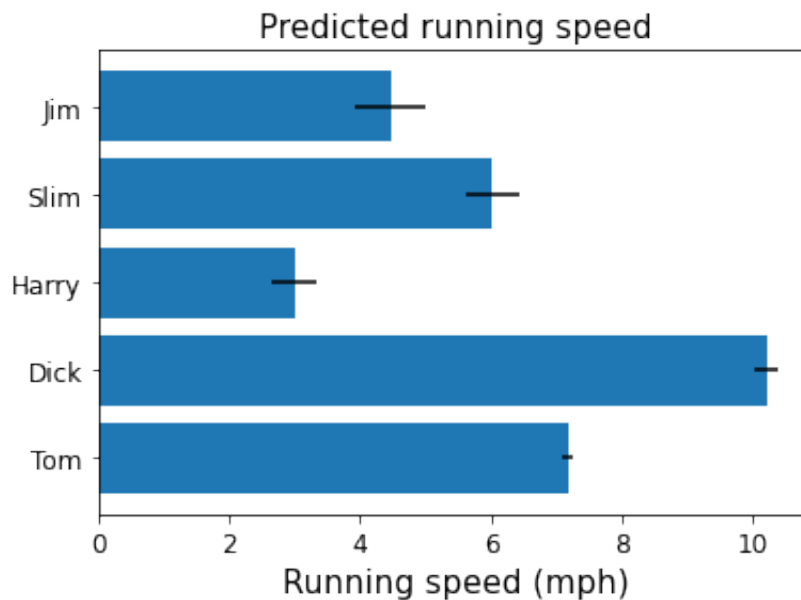
# Fixing random state for reproducibility
np.random.seed(1)

fig, ax = plt.subplots()

# Example data
people = ('Tom', 'Dick', 'Harry', 'Slim', 'Jim')
y_pos = np.arange(len(people))
performance = 3 + 10 * np.random.rand(len(people))
std = np.random.rand(len(people))

ax.barh(y_pos, performance, xerr=std, align='center')
plt.yticks(y_pos, labels = people, size = 12)
plt.xticks(size = 12)
ax.set_xlabel('Running speed (mph)', size = 15)
ax.set_title('Predicted running speed', size = 15)

plt.show()
```



2.2 Prediction of class needs uncertainty estimates

Suppose we are given the temperature data for past 1 month. Using this data we need to predict whether the next day would be hot or cold.

For problems like these, just predicting hot/cold is not very useful as we don't convey how certain we are about the prediction. So we need to associate the prediction with a probability number to make the prediction more useful. For example: **We are 90% certain it would be hot tomorrow**

2.3 Machine learning algorithms we discuss are developed using probability and statistics concepts

The algorithms such as Linear Regression, Logistic Regression and other probabilistic version of standard algorithms that we discuss have been developed using the foundations of probability and statistical inference theory. Hence for properly understanding these algorithms, we also need to spend time on understanding these involved concepts

2.4 Machine learning models are trained using Probability concepts

The concepts such as:

1. Maximum Likelihood Estimation (used in Linear Regression)
2. Maximum a Posteriori Estimation (used in Probabilistic Regression)
3. Expectation Maximization (used in Clustering)..... and so on

form the building block for training most data mining and machine learning approaches. These concepts are directly/indirectly derived from mathematical theory of Probability, Statistics and Linear Algebra.

2.5 Understanding the scope of machine learning models

Not all models are applicable for all kind of datasets. Understanding statistics and probability theory will enable us to understand the validity of a model for a particular given dataset. In this regard **statistical diagnostic** methods have been found to be really helpful. These include methods like:

1. R^2
2. Hypothesis testing
3. F-Test... and so on

References

- In []: