Machine Learning

Regression

Example

• Supervised regression learning
• Build a model predicting how much it rains based on barometric pressure
• General trend of as barometric pressure decreases, we typically have more rain and as it increases we have less rain
• Goal: create a model predicting how much it will rain.

• Classic Solution:
  – Fit a line to the data
  – Linear regression → done th
  – X is barometric pressure
  – m and b are our parameters (enabling us to predict rain for any new input).
• Evaluation:
  – Decent, but is inaccurate in a few regions, where line does not really correspond to our data.
• What do do?

• Make a more complex model
• Fit a polynomial, add terms to fit data better
• Later: throw away data, use equation to predict rain for ‘new’ data.
Approach 2: K nearest Neighbor

- data-centric approach or instance based approach
- \( X = \text{down by 5 mm} \)
- Want to know how much it will rain.
- \( K = 3 \).

How should we use the points

a) Use the averages of their x-values
b) Take the largest Y
c) Use the mean of the Y values

‘Building’ the model

- Repeating process at many points along the X-axis
- Curves ‘naturally’
- Other methods that keep data around: example: kernel regression.
  - Weigh contributions of data points.
  - KNN – neighbors have equal weight
**Best model to use: Parametric of Non?**

- **Example 1**: How far will the cannon ball go depending on what is \( x \) (angle of at which the barrel is aimed)

- **Example 2**: How many honeybees will visit a food source as we change the richness of the food source from, say, poor to rich.

**Example 1**: The cannon ball distance can be best estimated using a parametric model, as it follows a well-defined trajectory.

**Example 2**: Behavior of honey bees can be hard to model mathematically. Therefore, a non-parametric approach would be more suitable.

---

**Training & Testing**

- \( X_1, X_2 \), different features
- \( Y \) is what we are predicting.
- How to evaluate our learning algorithm?
  - Divide it up into Training and Testing.

- The more closely the model outputs a \( Y \) that reflects this \( X_{\text{test}} \) data, the more accurate the model is.
- **Split**: Time oriented
Decision Tree Board.

- List of factors $X_1, X_2, X_n$
- Labels (leaf nodes)