349:Machine Learning

Fall 2024

Final Projects
and
Selected Machine Learning Techniques

Final Projects -- Proposals

Proposal Elements:

1. Task description (define task, motivation optional)
2. Dataset (where will you acquire data? how will you create data?)
3. Features/attributes (how will you select and construct your features?)
4. Project execution (what steps will you take to complete the project?)
   - data preprocessing and handling
   - machine learning techniques (baseline(s) and a neural network)
   - interpretation and analysis of results

   Note: project must include a quantitative evaluation metric

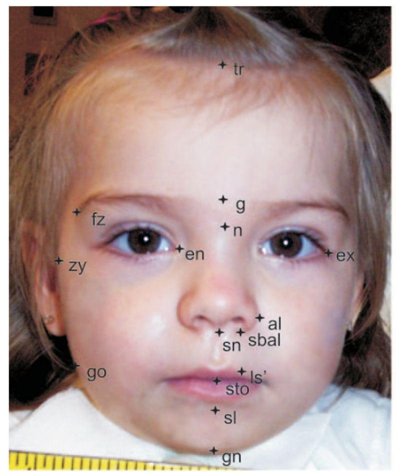
Other Considerations:

1. Scope project to be “doable” within the quarter
2. Guidance on discussion and analysis of results will come later

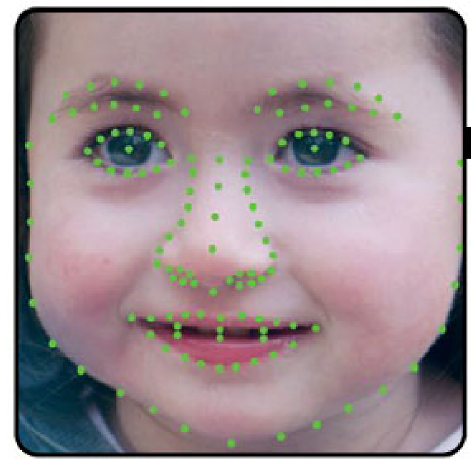
Final Projects -- Potential Ideas

Diagnostic Screening Using Facial Features:

vs.



vs.

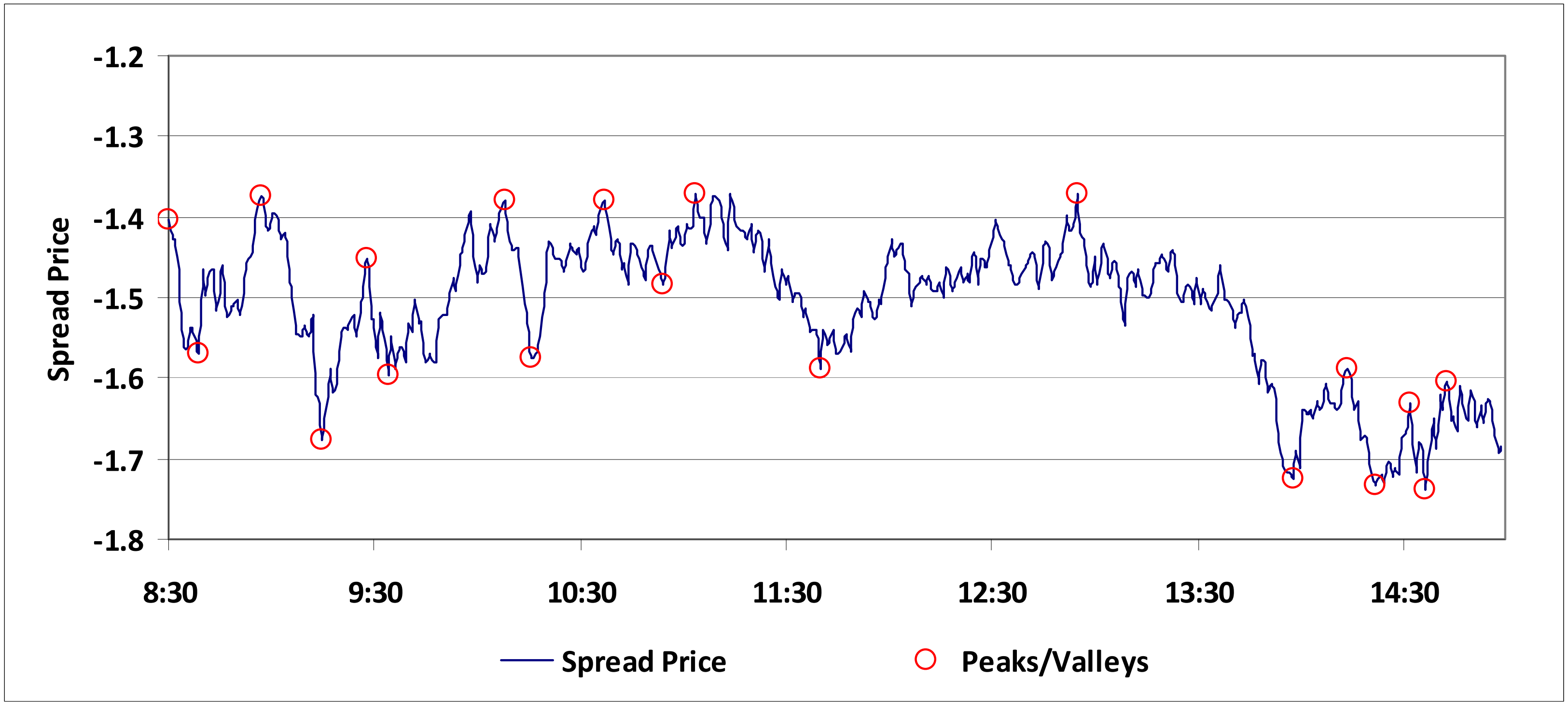


From: Identifying facial phenotypes of genetic disorders using deep learning, (Gurovich et al 2019), and

Facial Phenotype in Children and Young Adults (Todd et al 2005).

Final Projects -- Potential Ideas (*Cont.*)

Stock Price Prediction:



Final Projects -- Potential Ideas (*Cont.)*

Previous Sections of 349:

- https://users.cs.northwestern.edu/~ddowney/courses/349_Fall2008/projects/
- https://users.cs.northwestern.edu/~ddowney/courses/349_Fall2010/projects.html
- https://users.cs.northwestern.edu/~ddowney/courses/349_Winter2014/projects.html
- https://users.cs.northwestern.edu/~ddowney/courses/349_Spring2015/completed_projects.html

Kaggle Projects

When Should We Use Machine Learning?

- Well defined problem
- Lack of "easy" solution
- Large amounts of high quality data
- Clear and meaningful evaluation
- When the solution is justified

Well Defined Problem

- Machine learning methods optimize some mathematical goal
- Fuzzy problem definitions do not work
- Problem should have a clear and formal definition
- Questions to ask:
  - Can you formally write out the problem statement(s)?
  - Can people agree on how the system should behave?

Lack of “Easy” Solutions

- Lack of deterministic, rule-based solutions
  - ML is fundamentally different from “traditional” software
  - Most software has written specs that precisely describe system behavior
  - ML is inherently unpredictable as it changes with input
- What kind of problem do you have?
  - Not ML: You can hardcode an expert system to solve the problem
  - ML: Can provide examples, but cannot easily code solution
  - ML: Uses probabilistic reasoning

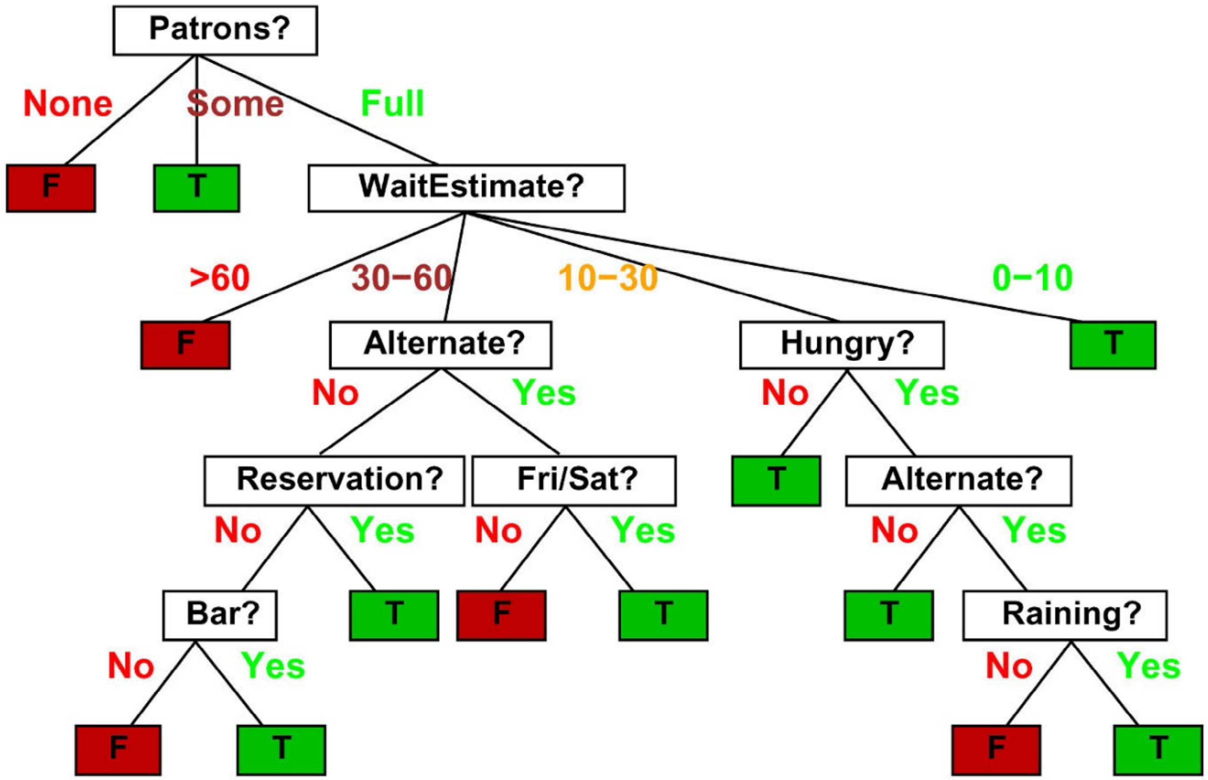
Large Amounts of High Quality Data

- ML success depends entirely upon high quality data
- What is **good** data?
  - o Large enough: do you have enough examples?
  - o Representative: do collected data correctly reflect the problem?
  - o Quality annotations: are the labels clear, consistent replicable, in sufficient quantity and cost effective?

Clear and Meaningful Evaluation

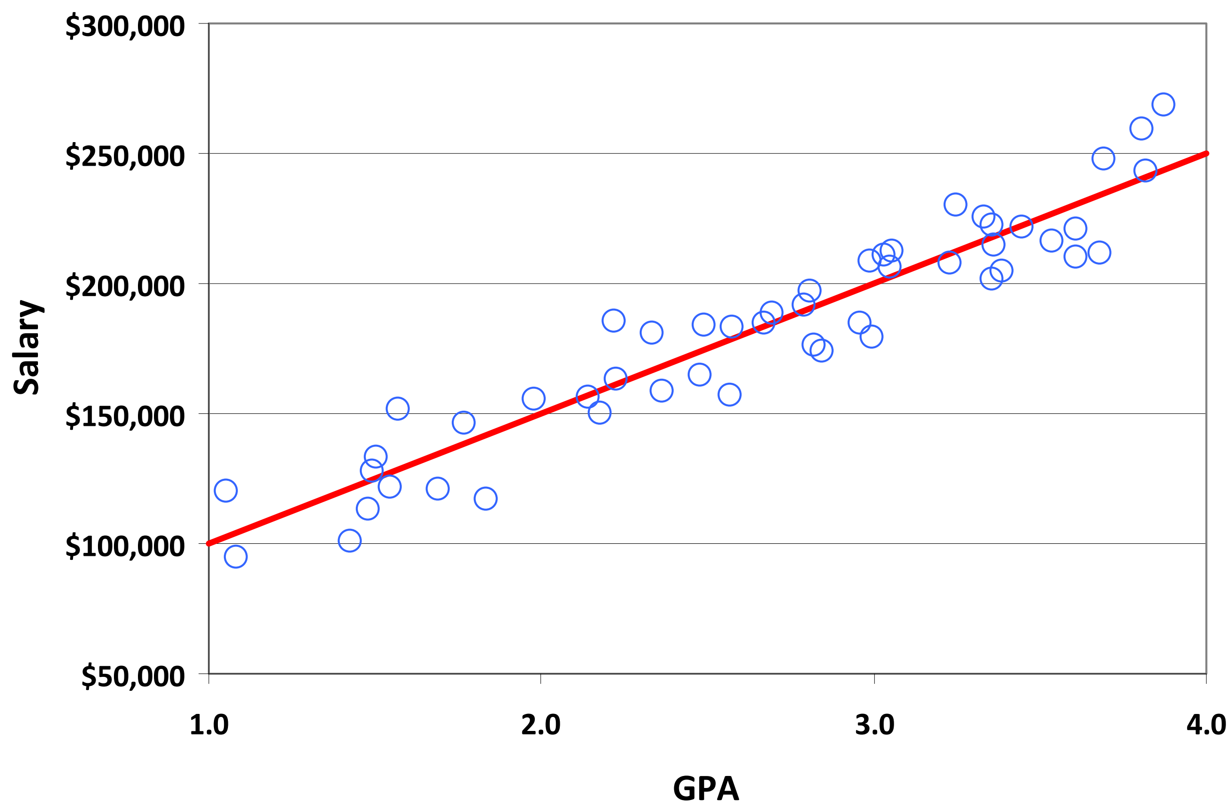
- Can you evaluate how well you are doing?
- Some goals are very difficult to measure
  - Not easily observed, expensive to obtain, too rare
- Can we measure what we actually care about?
  - Often times we use easier-to-measure proxies
- Can we quantify cost of different mistakes?
- Can we understand systematic biases and edge cases?

Machine Learning Techniques -- Decision Trees



$$I(X,Y) = H(X) - H(X \mid Y) \qquad H(X) = \sum_{i}^{n} -P(x=i) \cdot \log_2 P(x=i)$$

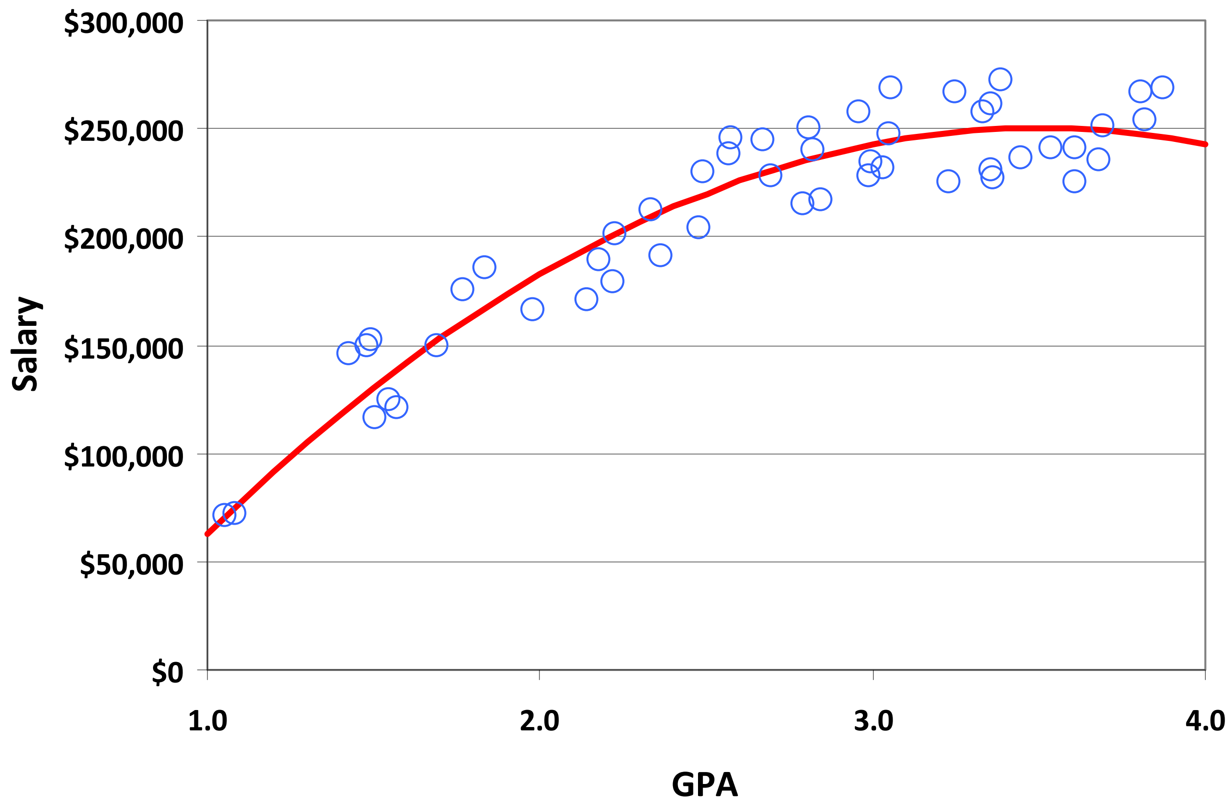
Machine Learning Techniques -- Linear Regression



$$h(x) = w_0 + w_1 x_1$$

Note: Data is randomly constructed purely illustrative purposes.

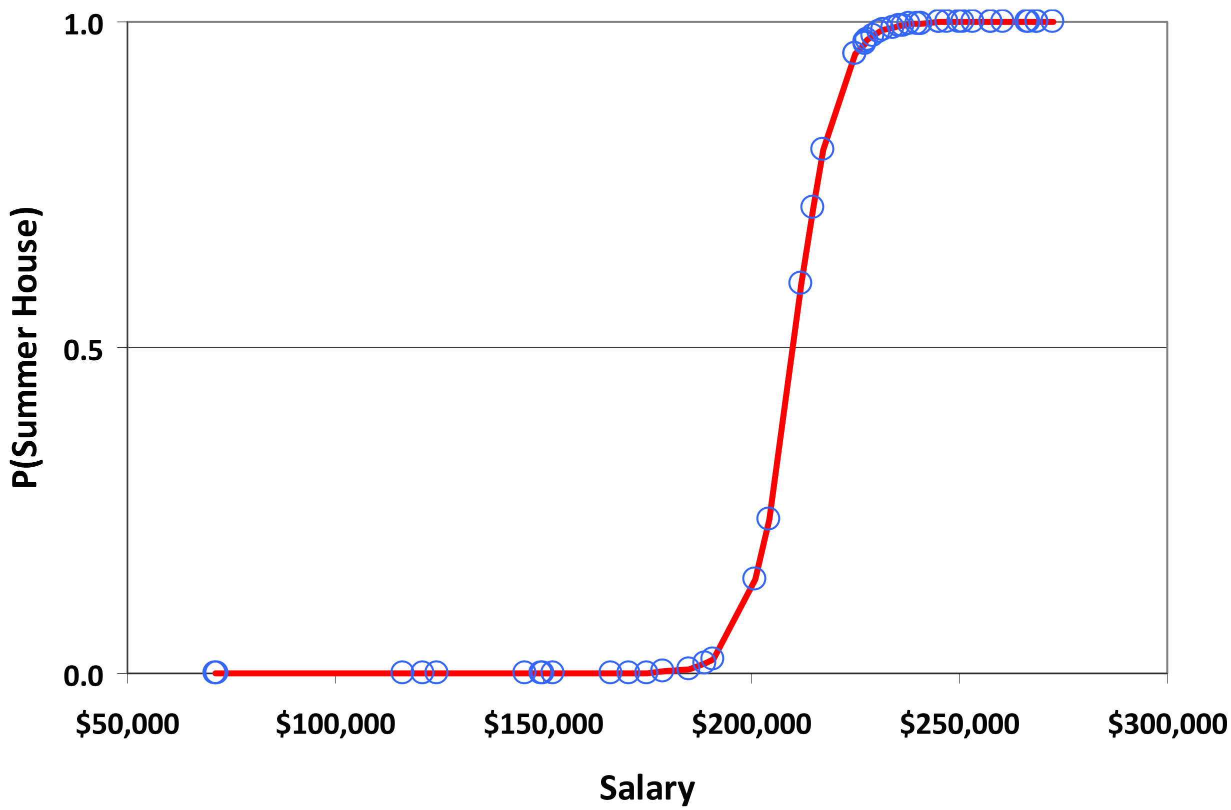
Machine Learning Techniques -- Polynomial Regression



$$h(x) = w_0 + w_1 (x_1 - x_0)^{k_1}$$

Note: Data is randomly constructed purely illustrative purposes.

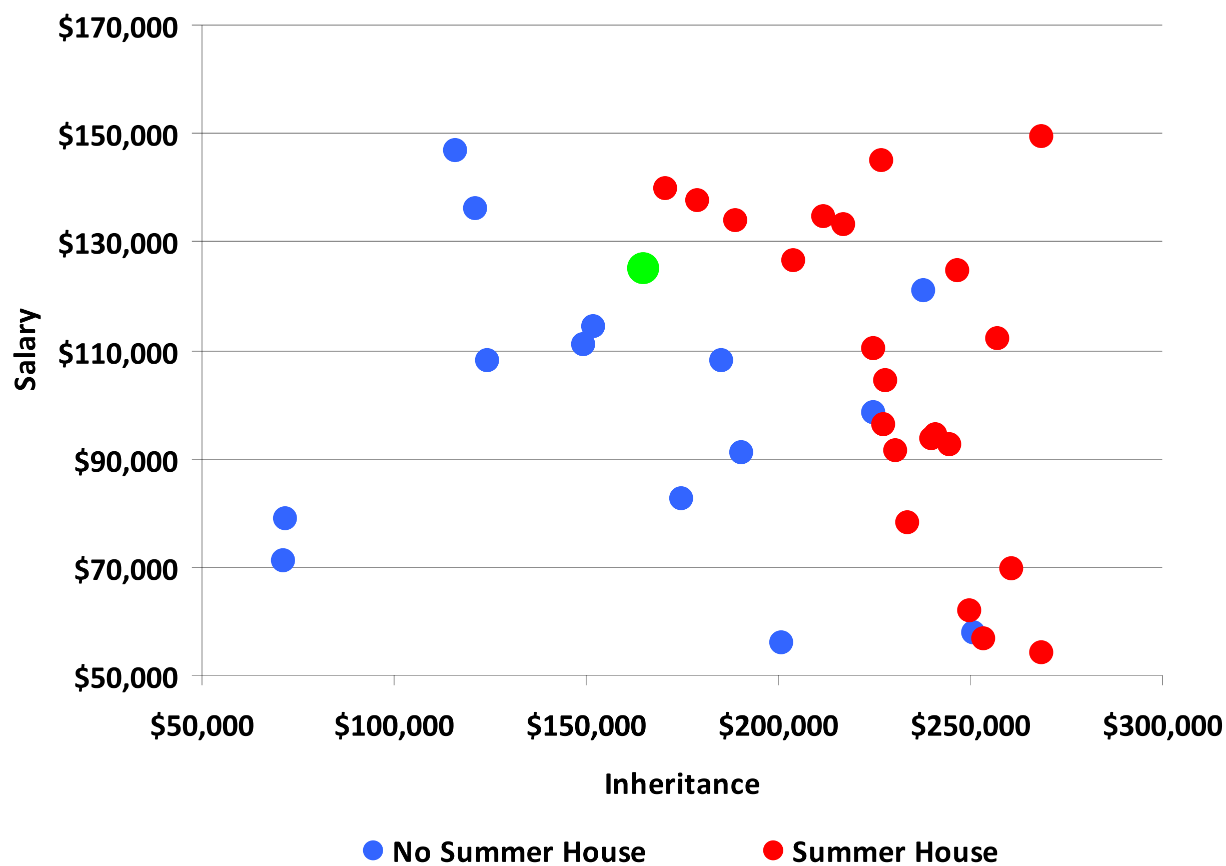
Machine Learning Techniques -- Logistic Regression



$$h(x) = \sigma(F(x)) = \frac{1}{1 + e^{-F(x)}}$$

Note: Data is randomly constructed purely illustrative purposes.

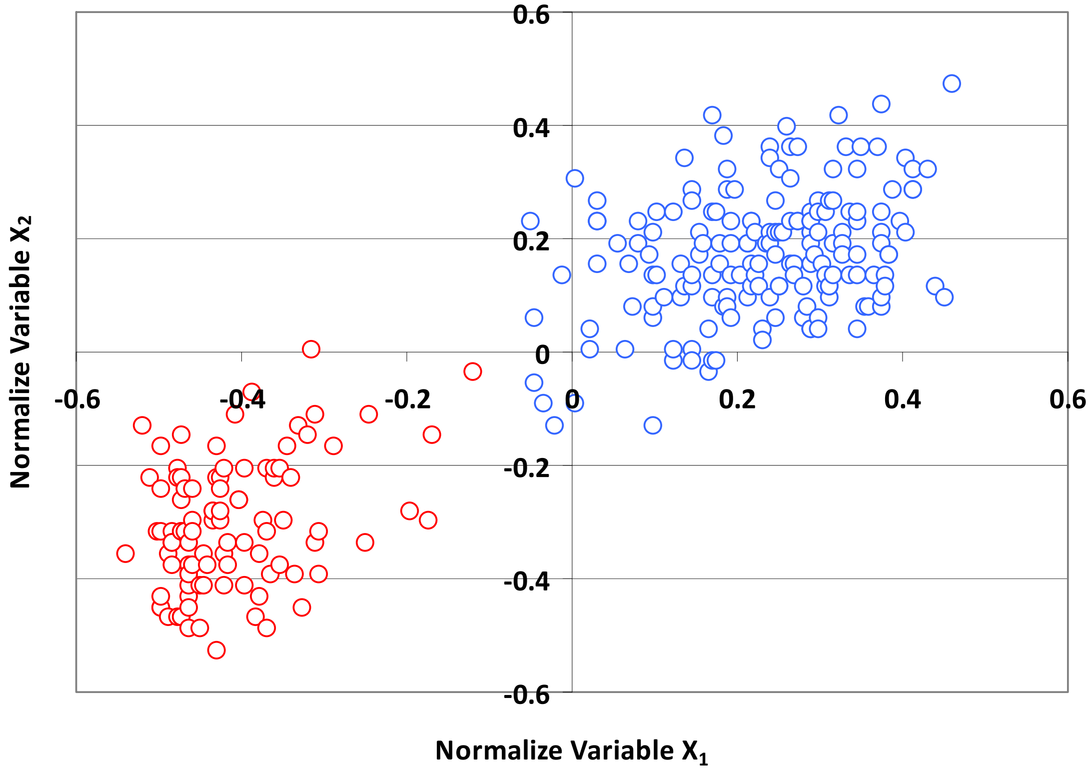
Machine Learning Techniques -- K-Nearest Neighbors



$$\vec{x}_{nn} = \arg\min\left(d\left(\vec{x}, \vec{x}_q\right)\right)$$

Note: Data is randomly constructed purely illustrative purposes.

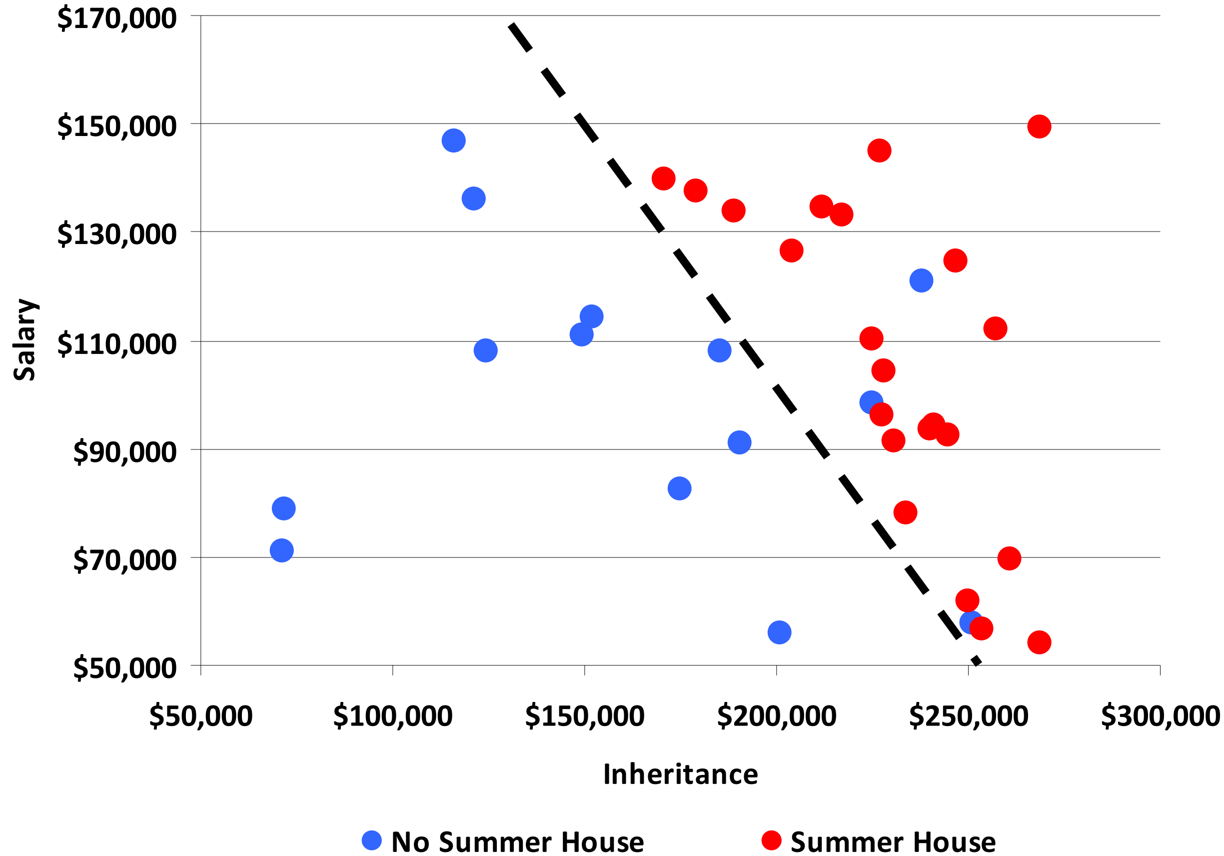
Machine Learning Techniques -- K-Means Clustering



$$c_i \equiv \arg\min_m \left\| x_i - \mu_m \right\|_2 \qquad \mu_{m,a} \equiv \frac{\sum_i \mathbf{1}_{c \in m} \cdot x_{i,a}}{\sum_i \mathbf{1}_{c \in m}}$$

Note: Data is randomly constructed purely illustrative purposes.

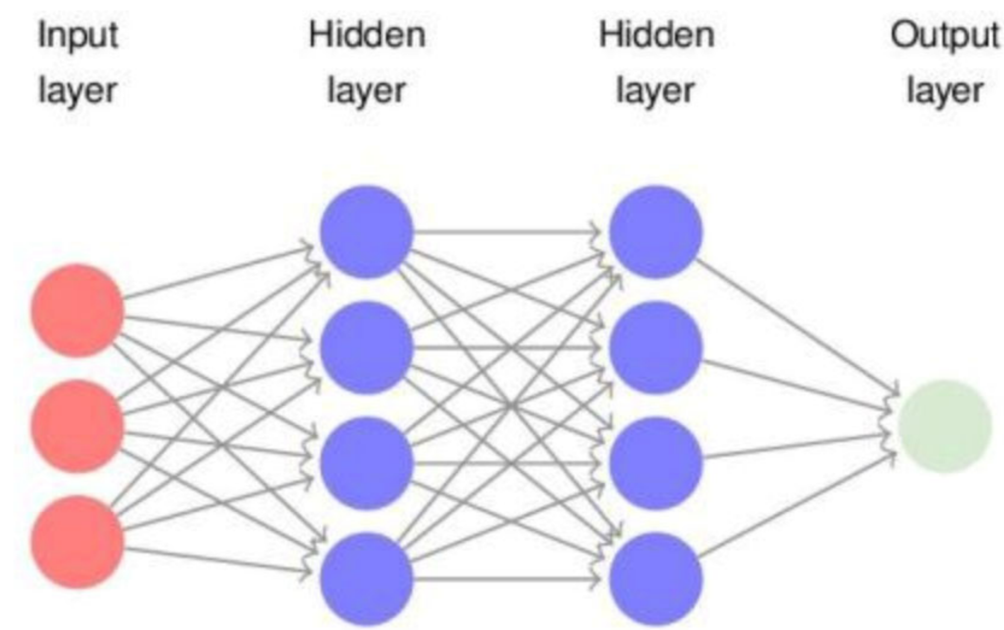
Machine Learning Techniques -- Linear Discriminants



$$g(x) = w_0 + w_1 x_1 + w_2 x_2 = 0 \qquad h(x) = \begin{cases} 1 & \texttt{if } g(x) > 0 \\ -1 & \texttt{otherwise} \end{cases}$$

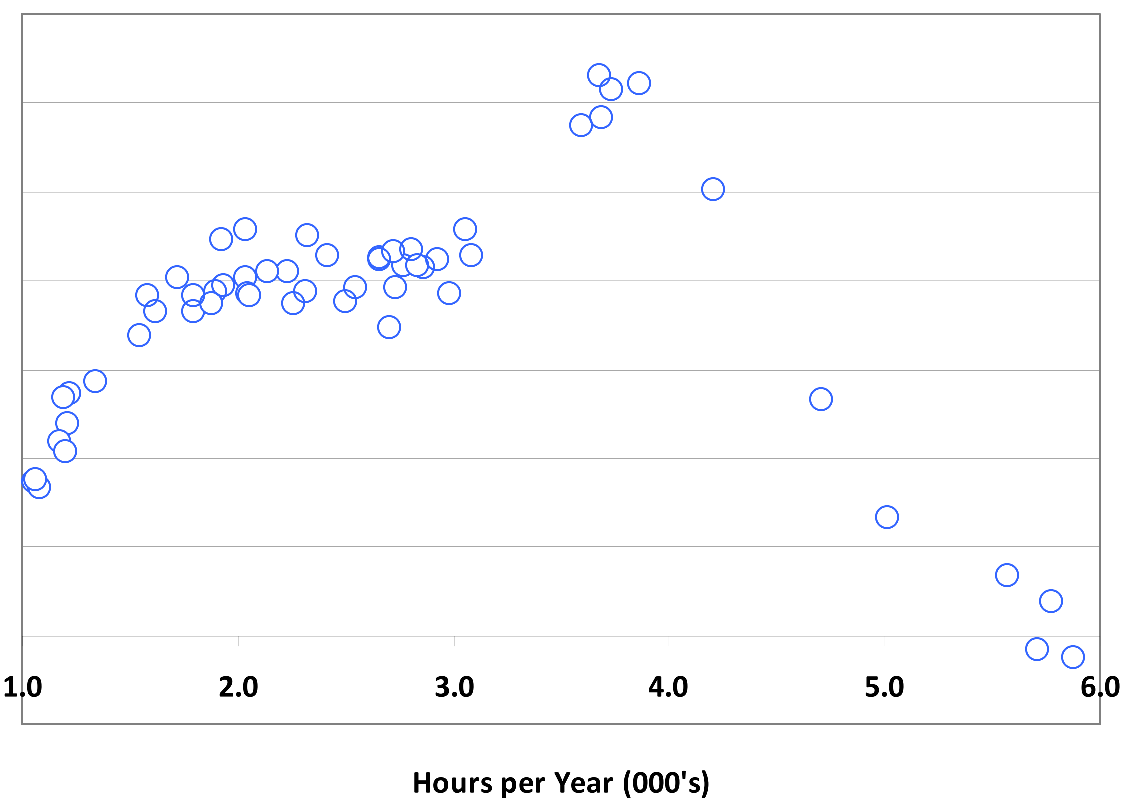
Note: Data is randomly constructed purely illustrative purposes.

Machine Learning Techniques -- Neural Networks

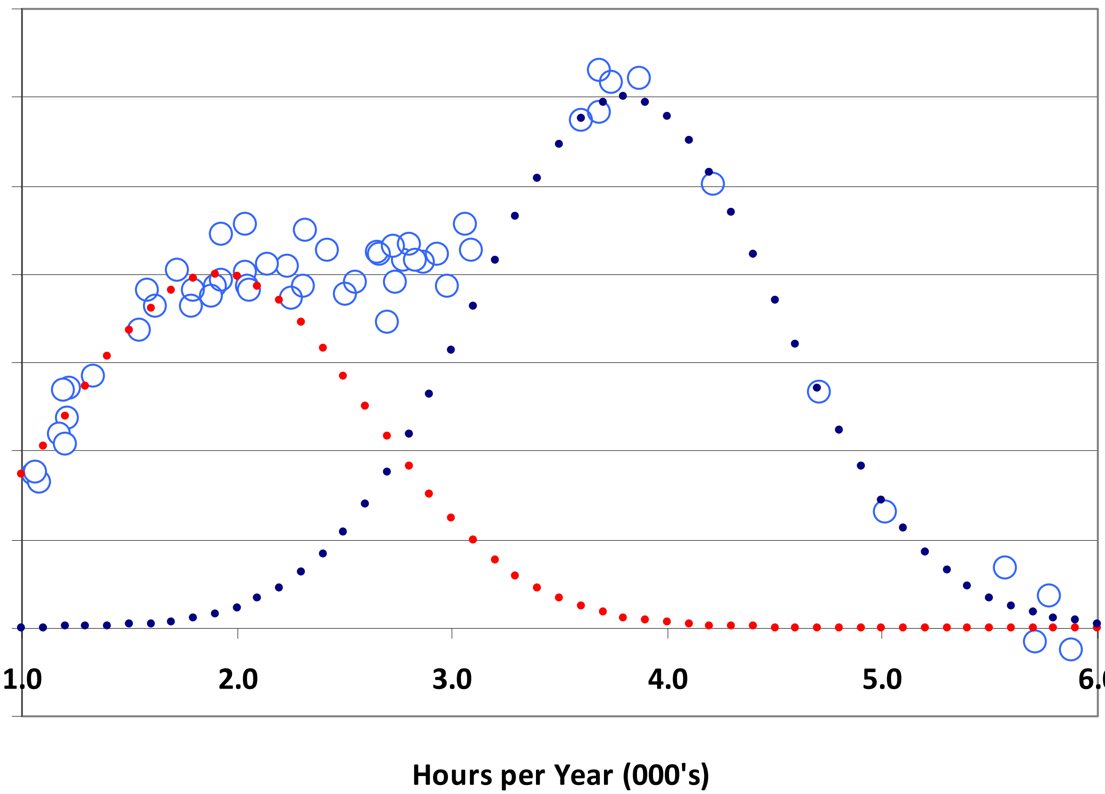


$$\hat{y}_k = g\left(\sum_{j=0}^{M} w_{jk} f\left(\sum_{i=0}^{N} w_{ij} x_i + b_j\right) + b_k\right)$$

Machine Learning Techniques -- Gaussian Mixture Model

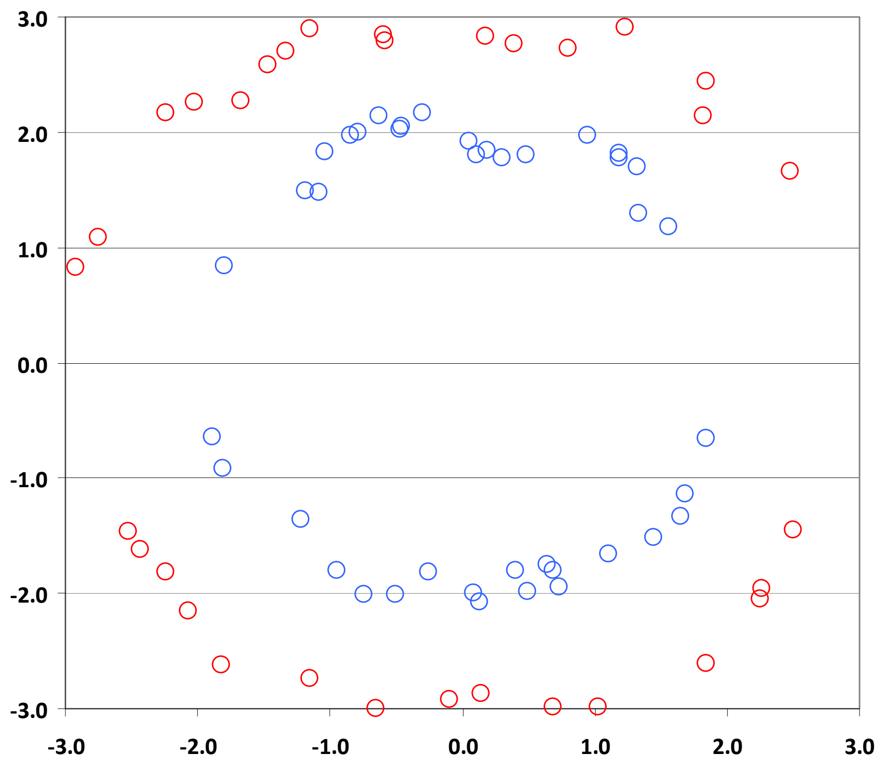


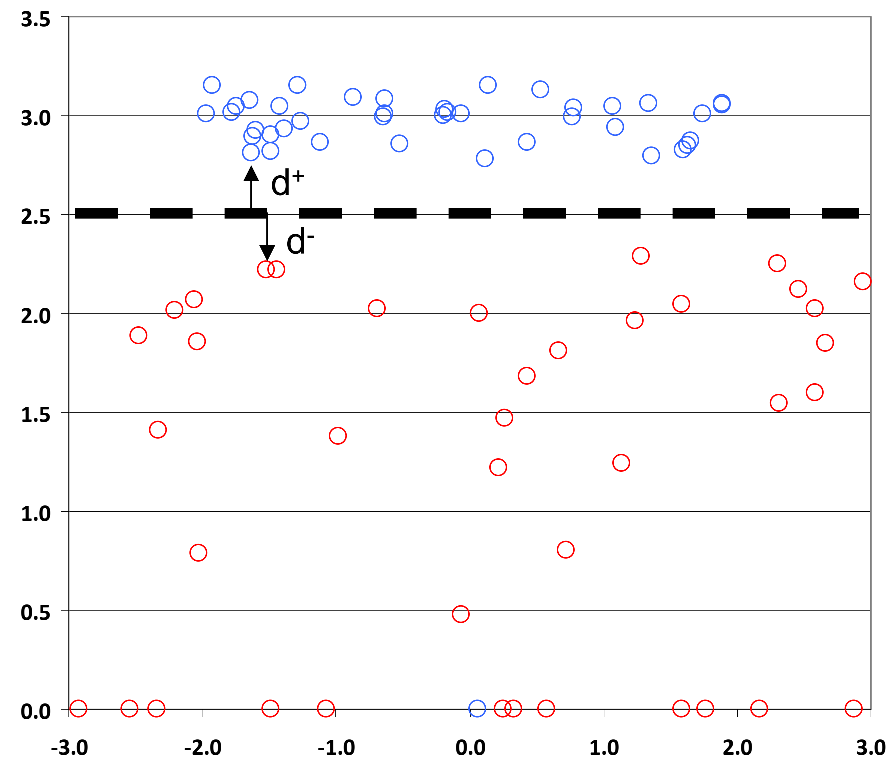
Machine Learning Techniques -- Gaussian Mixture Model



$$P(x) = \sum_{k=1}^{K} w_k N(x \mid \mu_k, \sigma_k^2) \qquad \sum_{k=1}^{K} w_k = 1$$

Machine Learning Techniques -- Support Vector Machine





$$x \rightarrow \phi(x) \qquad d^{+} = d^{-} = \frac{1}{\|w\|}$$

$$\max \frac{2}{\|w\|} \quad \texttt{such that} \quad y_i(w \cdot x_i + b) \geq 1, \forall \{x_i, y_i\} \in D$$

Regression → Classification

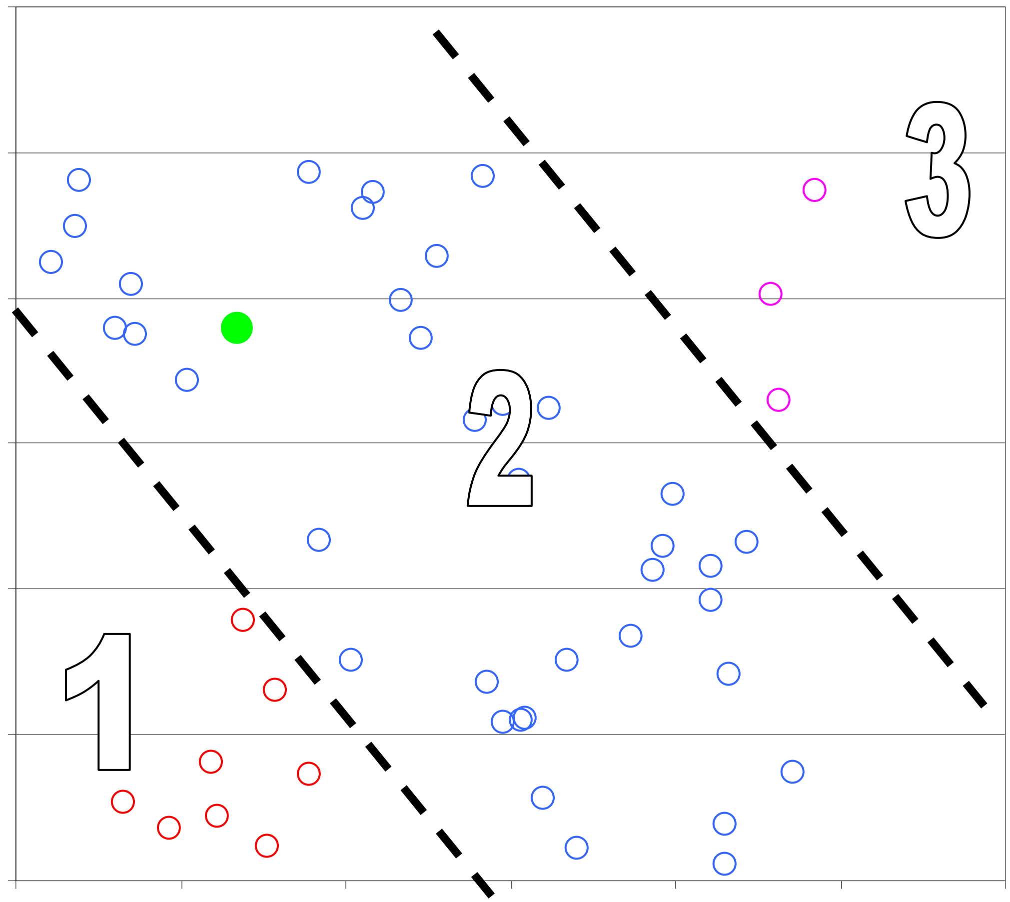
The *softmax* function maps *logits* to probability a distribution:

$$P_{C=c_i}(\hat{y}_i)) = \frac{e^{y_i^T w}}{\sum_C e^{y_c^T w}}$$

such that:

$$0 \le P_c \le 1 \qquad\qquad \sum_C P_c = 1.0$$

Classification → Regression



● ≈ 1.3