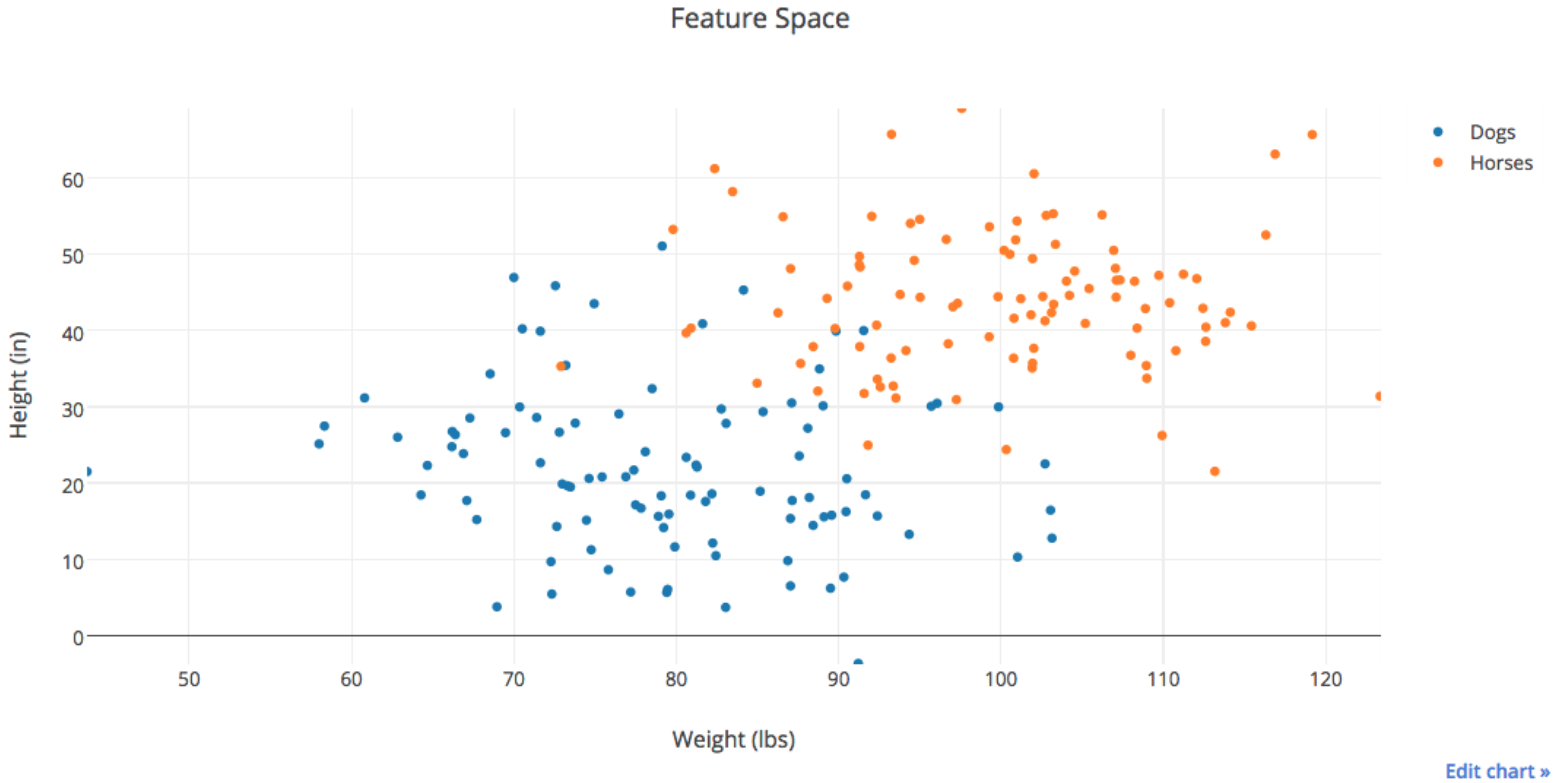
K Nearest Neighbors

How would you classify a new observation?

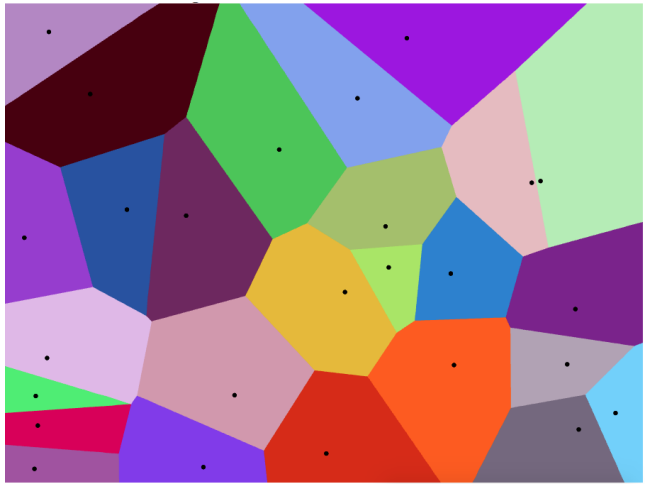


K-NN Algorithm Summary

- Find k nearest nearest neighbors to point of interest
- Count how many of those k neighbors are of each class
- Classify the point of interest as the majority class

k-NN Decision Boundaries

See IPython Notebook



Distance Metrics

Euclidean Distance:

$$\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} = \|\vec{x}\|_2$$

Distance Metrics

Cosine Similarity:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Distance Metrics

- Most common choices are Euclidean (L^2 norm) and cosine similarity
- Many others to choose from:
  - Manhattan distance (L^1 norm)
  - L^p norm
  - L^infinity norm
  - Hamming distance

k-NN Details

- Instance-based learning
- “Lazy” learning
- Sometimes called the first machine learning method (invented in 1950s)

How to Choose k?

- Cross-validation
- Rule of thumb: start with k = sqrt(n)

Pros/Cons of k-NN

- **Pros**
  - works with any number of classes
  - easy to store the model
    (it's just the data plus your distance metric)
  - can learn a very complex function
- **Cons**
  - slow
  - irrelevant attributes can affect results
  - curse of dimensionality

Uses of k-NN

- Classification
- Imputation
  - Replace missing values with k-NN
    http://nerds.airbnb.com/overcoming-missing-values-in-a-rfc/
    http://www.icmc.usp.br/~gbatista/files/his2002.pdf
- Anomaly Detection
  - e.g. use distance to kth nearest neighbor is an outlier score

Variants of k-NN

- Weighted k-NN
- Edited knn

k-NN Theoretical Guarantees

Behavior in the Limit

$\epsilon^*(\mathbf{x})$: Error of optimal prediction
$\epsilon_{NN}(\mathbf{x})$: Error of nearest neighbor

Theorem: $\lim_{n\to\infty} \epsilon_{NN} \leq 2\epsilon^*$

Proof sketch (2-class case):

$$\begin{aligned}\epsilon_{NN} &= p_+ p_{NN\epsilon-} + p_- p_{NN\epsilon+} \\ &= p_+(1 - p_{NN\epsilon+}) + (1 - p_+)p_{NN\epsilon+}\end{aligned}$$

$\lim_{n\to\infty} p_{NN\epsilon+} = p_+, \quad \lim_{n\to\infty} p_{NN\epsilon-} = p_-$

$\lim_{n\to\infty} \epsilon_{NN} = p_+(1-p_+)+(1-p_+)p_+ = 2\epsilon^*(1-\epsilon^*) \leq 2\epsilon^*$

Theorem: $\lim_{n\to\infty,\, k\to\infty,\, k/n\to 0} \epsilon_{kNN} = \epsilon^*$

Slides by Pedro Domingos, Tom Mitchell and Thomas Dietterich