The nature of data unit:

- Objects and vectors (Wednesday)
- K nearest neighbors classification (**today**)
- (Start linear regression next week)

Today:

- Leftover: Meet the data; features as functions
- Concept
- Algorithm and analysis
- Things to notice
- Distance metrics and norms

Let X be the training data of N D-dimensional observations.

Let k be a natrual number.

Given a new datapoint $\vec{x}$, assign $\vec{x}$ t6he same class as the majority of k training datapoints that are closest to $\vec{x}$ in the vector space.

Intance-based methods can also use more complex, symbolic representations for instances. . . . Case-based reasoning has been applied to tasks such as storing and reusing past experience at a help desk, reasoning about about legal cases by referring to previous cases, and solving complex scheduling problems by reusing relevant portions of previously solved problems.

Mitchell, Machine Learning, p 231.

A *metric* or *distance function* between two vectors is a function $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ with the properties that for any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^D$,

- $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})$ (*symmetry*)
- $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ (*the triangle inequality*)
- $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$ (*the identity of indiscernibles*)

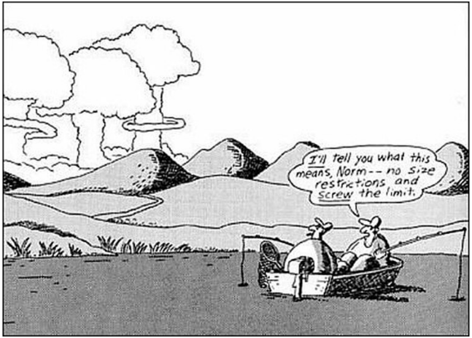
A *norm* is a fuction $||\cdot|| : \mathbb{R}^D \to \mathbb{R}$ with the properties that for any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ and any scalar $\lambda \in \mathbb{R}$,

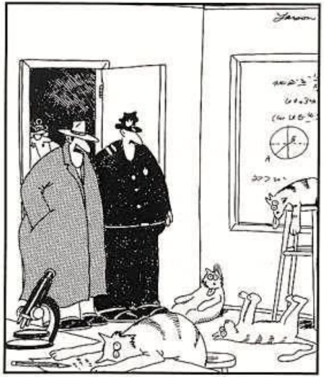
- $||\lambda \boldsymbol{x}|| = |\lambda| ||\boldsymbol{x}||$ (*absolute homogeneity*)
- $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$ (*the triangle inequality*)
- $||\boldsymbol{x}|| \geq 0$ and, moreover, $||\boldsymbol{x}|| = 0$ iff $\boldsymbol{x} = \boldsymbol{0}$ (*positive definiteness*)

Any norm induces a metric with

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||$$

Where the respective worlds of boating and herpetology converge.





"Notice all the computations, theoretical scribblings, and lab equipment, Norm. ... Yes, curiosity killed these cats."

Euclidean distance (L_2 norm):

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_0 - y_0)^2 + (x_1 - y_1)^2} \qquad d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\left(\sum_{i=0}^{D-1} (x_i - y_i)^2\right)}$$

Manhattan or city-block distance (L_1 norm):

$$d(\boldsymbol{x}, \boldsymbol{y}) = |x_0 - y_0| + |x_1 - y_1| \qquad d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{D-1} |x_i - y_i|$$

Minkowski distance (L_p norm):

$$d(\boldsymbol{x}, \boldsymbol{y}) = (|x_0 - y_0|^p + |x_1 - y_1|^p)^{\frac{1}{p}} \qquad d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=0}^{D-1} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Mahalanobis distance: Let $\mathbf{S}$ be the covariance matrix of the entire dataset $\mathbf{X}$, and so $\mathbf{S^{-1}}$ is the inverse of the covariance matrix.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \mathbf{S}^{-}1(\boldsymbol{x} - \boldsymbol{y})}$$

Compare with Euclidean distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\left(\sum_{i=0}^{D-1} (x_i - y_i)^2 \right)} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y})}$$

Canberra distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{D-1} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Compare with Manhattan distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = |x_0 - y_0| + |x_1 - y_1| \qquad d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{D-1} |x_i - y_i|$$

Coming up:

Due Fri, Jan 24:

Do numpy programming assignment

Take the KNN quiz

Due Wed, Jan 29:

Read and respond to article about the Boston Housing Dataset

(Find pdf on Canvas.)

Due Thurs, Jan 30:

Read the textbook from Chapters 1 and 5 (see Canvas for specific sections)

Due Fri, Jan 31:

Do KNN programming assignment

Due Mon, Feb 3:

Propose project topic