k-Nearest Neighbors

Introduction to Non-Parametric Learners

Cary Goltermann

Galvanize

2016

Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

Parametric Models

Models that can be described with a finite number of parameters.

- Linear Regression
- Logistic Regression
- Neural Networks

Non-parametric Models

Models whose parameters reside in an infinite-dimensioned parameter space.

- Decision Trees
- Support Vector Machines
- k-Nearest Neighbors

These models, variants of them, and their applications are what we're going to be learning about for the next two weeks.

Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

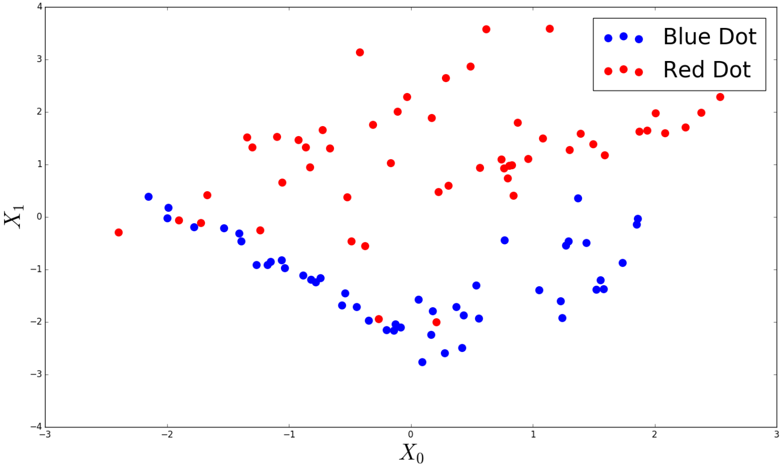
Curse of Dimensionality

- Completely Unintuitive

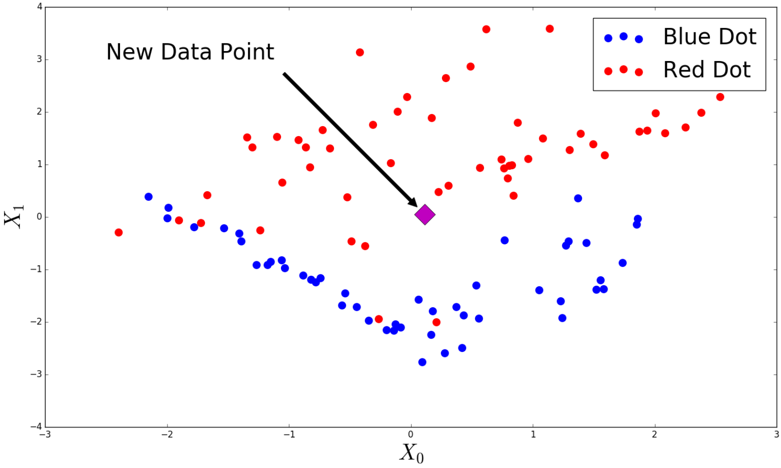
Intuition

- Consider trying to write your own supervised machine learning algorithm.
- It's expected receive data, X, and a corresponding targets, y, and with it be able to take it a new data point X_0 and output a prediction for its target.
- Let's stay simple for a moment and talk about ways that we could make such predictions.

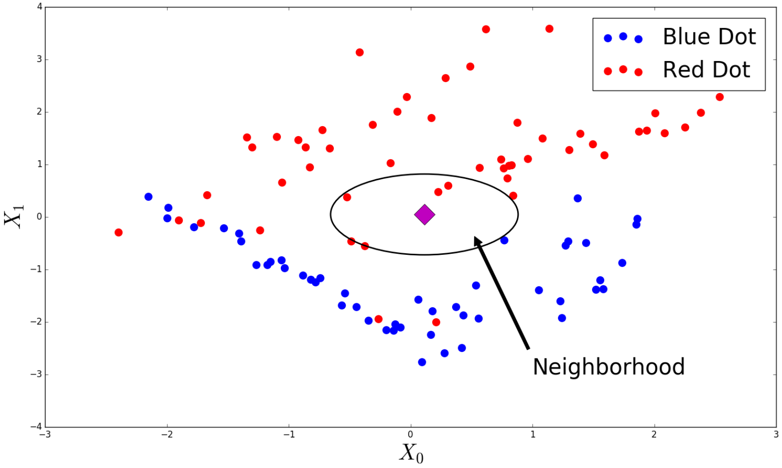
Example Data to Point At



New Data - How Do We Predict?



Check Out Some Close Points. They're Probably Similar!



Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

Training

Training

Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

Prediction

1. Calculate the distance between new points and all points in the training data.

Prediction

1. Calculate the distance between new points and all points in the training data.
2. Look at the k points nearest the new point.

Prediction

1. Calculate the distance between new points and all points in the training data.
2. Look at the k points nearest the new point.
3. Predict the modal class amongst those k points for classification. For regression, choose the average (potentially weighted by distance) of the k points.

Overview

Modeling

- Parametric vs. Non-parametric

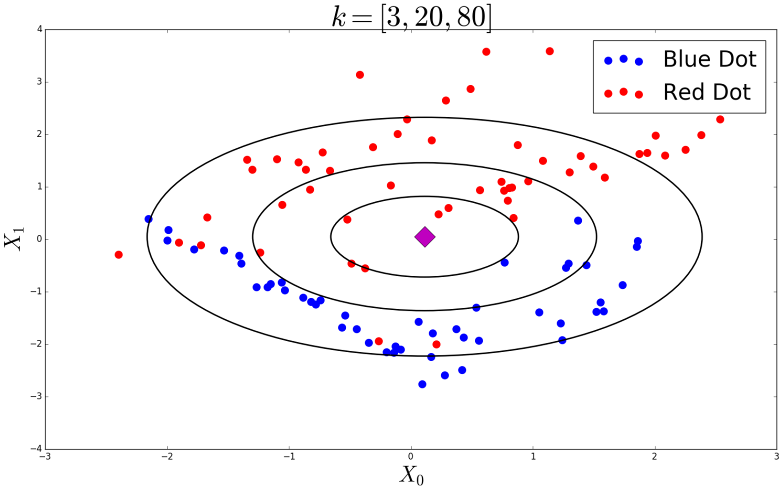
k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

How to Choose k



Bias-Variance Trade-off?

Question

How does changing k alter the bias and variance in our kNN models?

Bias-Variance Trade-off?

Question

How does changing k alter the bias and variance in our kNN models?

What happens when $k = 1$?

Bias-Variance Trade-off?

Question

How does changing k alter the bias and variance in our kNN models?

What happens when $k = 1$?

$\rightarrow$ High variance, low bias.

Bias-Variance Trade-off?

Question

How does changing k alter the bias and variance in our kNN models?

What happens when $k = 1$?

$\rightarrow$ High variance, low bias.

What happens when $k = n$?

Bias-Variance Trade-off?

Question

How does changing k alter the bias and variance in our kNN models?

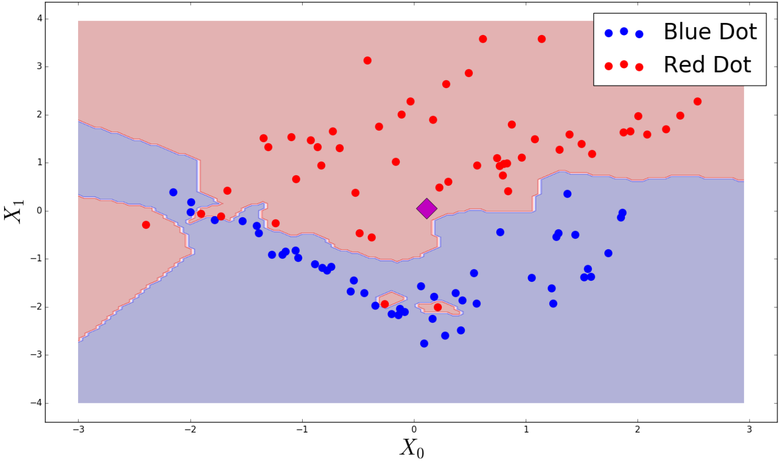
What happens when $k = 1$?

$\rightarrow$ High variance, low bias.

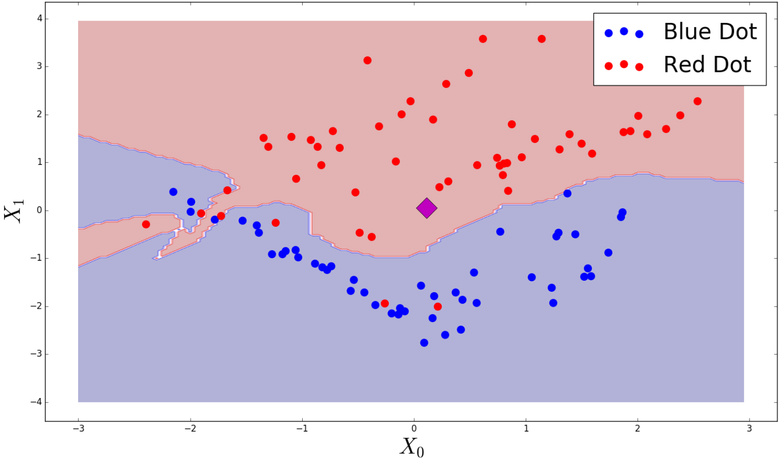
What happens when $k = n$?

$\rightarrow$ High bias, low variance.

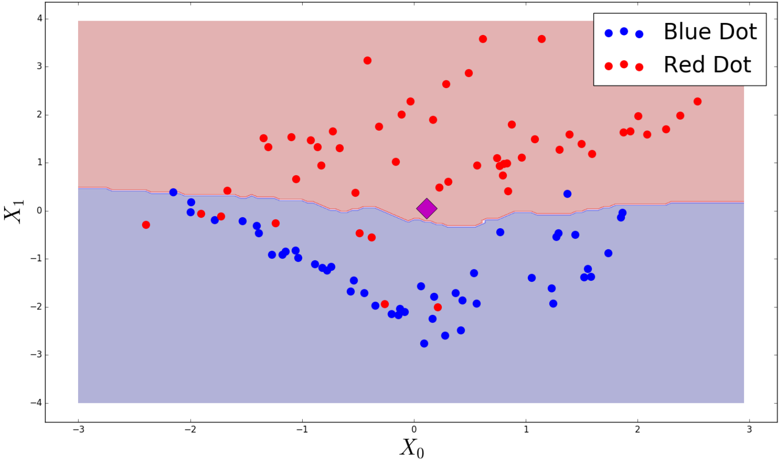
$k = 1$



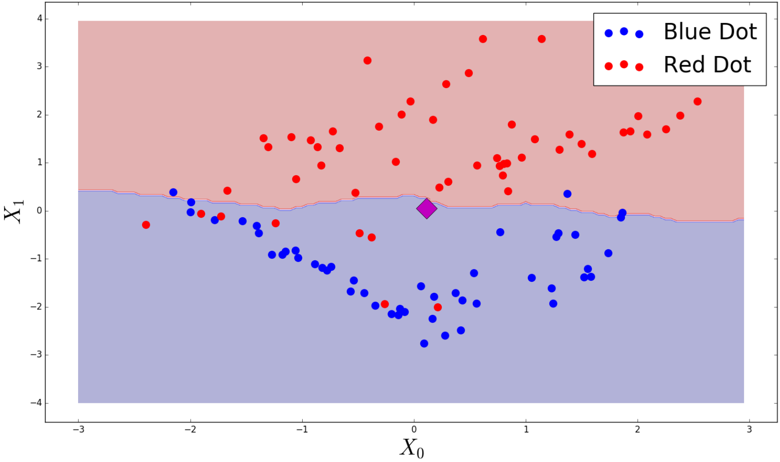
$k = 3$



$k = 20$



$k = 80$



Distance Metrics

So far we haven't been explicit about what distance metrics we're using. Some choices:

Euclidean $\sqrt{\sum_i (a_i - b_i)^2}$

Manhattan $\sum_i |a_i - b_i|$

Cosine $1 - \frac{a \cdot b}{||a||||b||}$

Pros

- Simple to use and intuitive.
- Extremely fast to train ;)
- Doesn't assume anything about the relationship between Xs and y.

Pros and Cons

Pros

- Simple to use and intuitive.
- Extremely fast to train ;)
- Doesn't assume anything about the relationship between Xs and y.

Cons

- Slow to predict.
- Unreliable with high dimensional data.
- Doesn't work well with categorical features.

Advice

- A good place to start searching for k is $\sqrt{n}$.
- Make sure to normalize data before training.
- Not a widely used algorithm for the reasons we saw above, some potential uses:
  - Missing data imputation.
  - Anomaly detection.

Variants

- One variant is to weight the votes by $\frac{1}{d_i}$ so closer points get more weight.
- Use for regression, take (optionally, weighted) mean of continuous target rather than vote.
- Approximate nearest neighbors, overcomes performance issues.

Overview

Modeling

- Parametric vs. Non-parametric

k-Nearest Neighbors

- Intuition
- Training
- Prediction
- Choosing k

Curse of Dimensionality

- Completely Unintuitive

Our Intuition is Lacking

The intuition that we have developed as beings living in a spatially 3-dimensional world fails us when we try to reason about how far apart things are in high dimensions.

Hypercubes

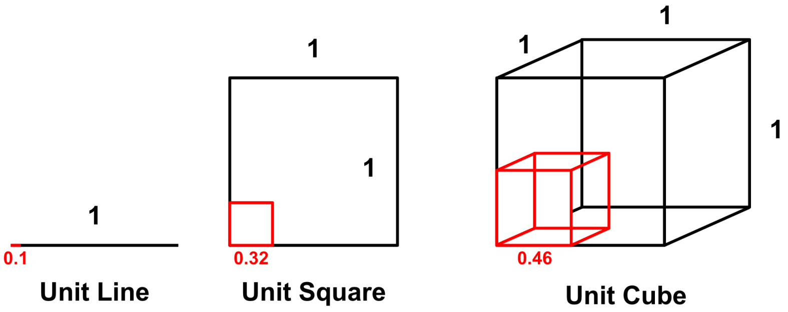
Question

Consider a hypercube inside a unit-hypercube. How long would we need to make the side of the smaller hypercube to capture 10% of the volume of the unit-hypercube?

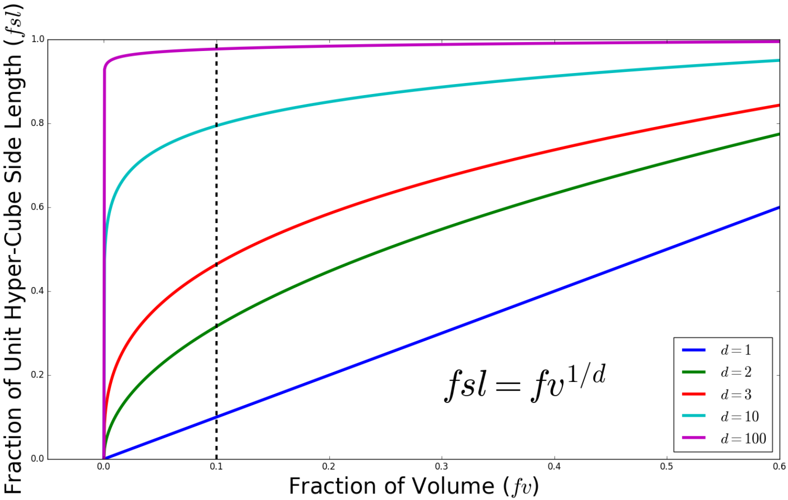
Hypercubes

Question

Consider a hypercube inside a unit-hypercube. How long would we need to make the side of the smaller hypercube to capture 10% of the volume of the unit-hypercube?



She's a Witch!



Fraction of Unit Hyper-Cube Side Length (fsl)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Fraction of Volume (fv)
$fsl = fv^{1/d}$
$d=1$
$d=2$
$d=3$
$d=10$
$d=100$