Introduction to Machine Learning for Social Science
Class 8: K Nearest Neighbor Algorithm

Rochelle Terman

Postdoctoral Fellow
Center for International Security Cooperation
Stanford University

February 1, 2018

Review

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

$$Y = f(X) + \epsilon$$

where f is fixed but unknown function of $X_1, ..., X_p$, and ϵ is a random error term that is is independent of X and has mean zero. .

Machine learning: estimating f with $\hat{f}$.

Review

Tools for estimating f:

- Linear probability regression

- Logit regression

- LASSO

⇝ parametric models

Parametric models

- Assumes function form of f
- Fixed parameters
- Estimate parameters from training data $(\hat{f})$
- Use $(\hat{f})$ for prediction
- Advantages: simplifies problem of estimating f to estimating parameters.
- Disadvantages: our assumptions might be wrong.

Non-Parametric models

- No explicit assumptions about function form of f
- No fixed parameters

K nearest neighbors
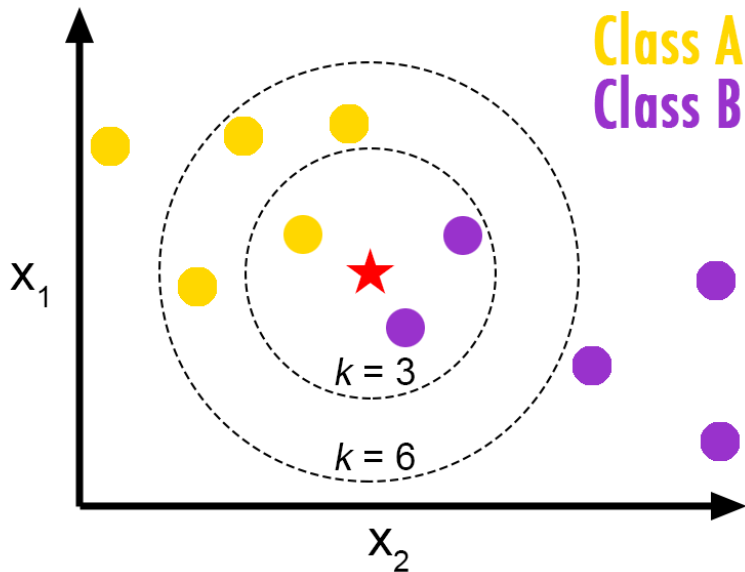
K nearest neighbors

- Among the simplest of all machine learning algorithms.

K nearest neighbors

- Among the simplest of all machine learning algorithms.
- Instance-based of "lazy" algorithm: function is only approximated locally and all computation is deferred until classification (no or very little training!).

K nearest neighbors

- Among the simplest of all machine learning algorithms.
- Instance-based of "lazy" algorithm: function is only approximated locally and all computation is deferred until classification (no or very little training!).
- Nonparametric! No fixed parameters! No explicit assumptions about function form of f!

K nearest neighbors

- Among the simplest of all machine learning algorithms.
- Instance-based of "lazy" algorithm: function is only approximated locally and all computation is deferred until classification (no or very little training!).
- Nonparametric! No fixed parameters! No explicit assumptions about function form of f!
- Choice of K has a drastic effect on the KNN classifier obtained: With $K = 1$, the decision boundary is overly flexible (low bias, high variance), while with $K = 100$ it is not sufficiently flexible (high bias, low variance).

Other nonparametric methods

- Decision Trees
- Support Vector Machines
- Neural Nets
- Random Forest

Questions for midterm?