IEOR165 Discussion
Week 7

Sheng Liu

University of California, Berkeley

Mar 3, 2016

Outline

Nadaraya-Watson Estimator: Motivation

Given the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we want to find the relationship between Y_i (response) and X_i (predictor):

$$Y_i = g(X_i) + \epsilon_i$$

Basically, we find to find $g(x_i)$, which is

$$g(x) = E(Y_i | X_i = x)$$

where we assume $E(\epsilon_i | X_i) = 0$.

- Parametric methods: assume the form of $g(x)$, e.g. linear, polynomial, exponential... Then we only need to estimate the parameters that characterize $g(x)$.
- Nonparametric methods: make no assumptions about the form of $g(x)$, which implies the number of parameters is infinite...

An Intuitive Nonparametric Method

K-nearest neighbor average:

$$\begin{aligned}\hat{g}(x) &= \frac{1}{k}\sum_{X_i \in N_k(x)} Y_i \\ &= \sum_{i=1}^{n} \frac{I(X_i \in N_k(x))}{k} \cdot Y_i \\ &= \sum_{i=1}^{n} \frac{I(X_i \in N_k(x))}{\sum_{i=1}^{n} I(X_i \in N_k(x))} \cdot Y_i \\ &= \frac{\sum_{i=1}^{n} I(X_i \in N_k(x)) \cdot Y_i}{\sum_{i=1}^{n} I(X_i \in N_k(x))}\end{aligned}$$

- You can compare this with our intuitive method for estimating pdf
- Not continuous and indifferentiable

Nadaraya-Watson Estimator

To reach smoothness, replace the indicator function by kernel function

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \cdot Y_i}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})}$$

Then we get the Nadaraya-Watson Estimator.

- Weighted average: kernel weights
- Relation to kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) K(\frac{Y_i - y}{h})$$

And by the formula

$$g(x) = E(Y|X = x) = \int y f(y|x) dy = \int y \frac{f(x, y)}{f(x)} dy$$

Nadaraya-Watson Estimator

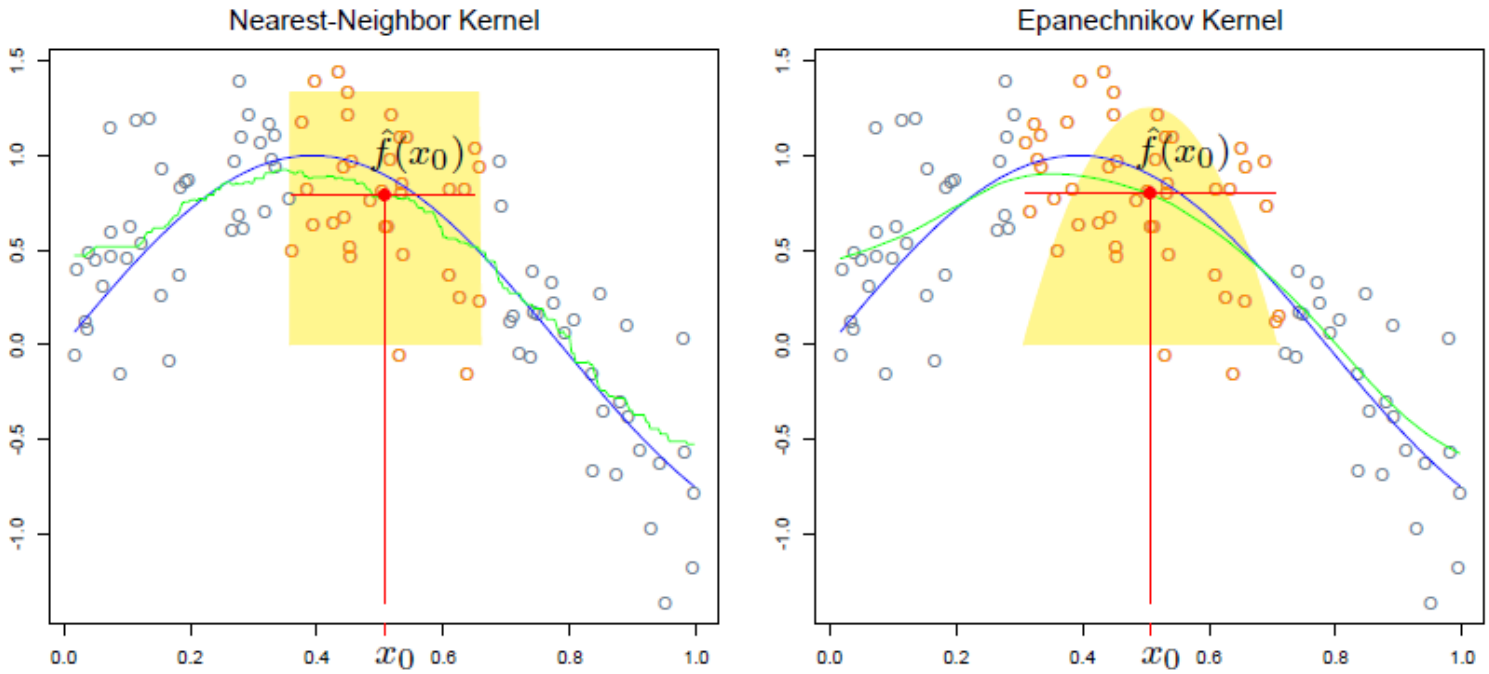


FIGURE 6.1. *In each panel* 100 *pairs* x_i, y_i *are generated at random from the blue curve with Gaussian errors:* $Y = \sin(4X) + \varepsilon$, $X \sim U[0, 1]$, $\varepsilon \sim N(0, 1/3)$. *In the left panel the green curve is the result of a* 30*-nearest-neighbor running-mean smoother. The red point is the fitted constant* $\hat{f}(x_0)$, *and the red circles indicate those observations contributing to the fit at* x_0. *The solid yellow region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width* $\lambda = 0.2$.

Partially Linear Model

A compromise between parametric model and nonparametric model

$$Y_i = X_i\beta + g(Z_i) + \epsilon_i$$

- Predictor: (X_i, Z_i)
- A parametric part: linear model
- A nonparametric part: $g(Z_i)$
- Estimation:

$$E(Y_i|Z_i) = E(X_i|Z_i)\beta + g(Z_i)$$

$$\Rightarrow Y_i - E(Y_i|Z_i) = (X_i - E(X_i|Z_i))\beta + \epsilon$$

- For $E(Y_i|Z_i)$ and $E(X_i|Z_i)$, use Nadaraya-Watson Estimators

Classification

When our response is qualitative, or categorical, predicting the response is actually doing classification.

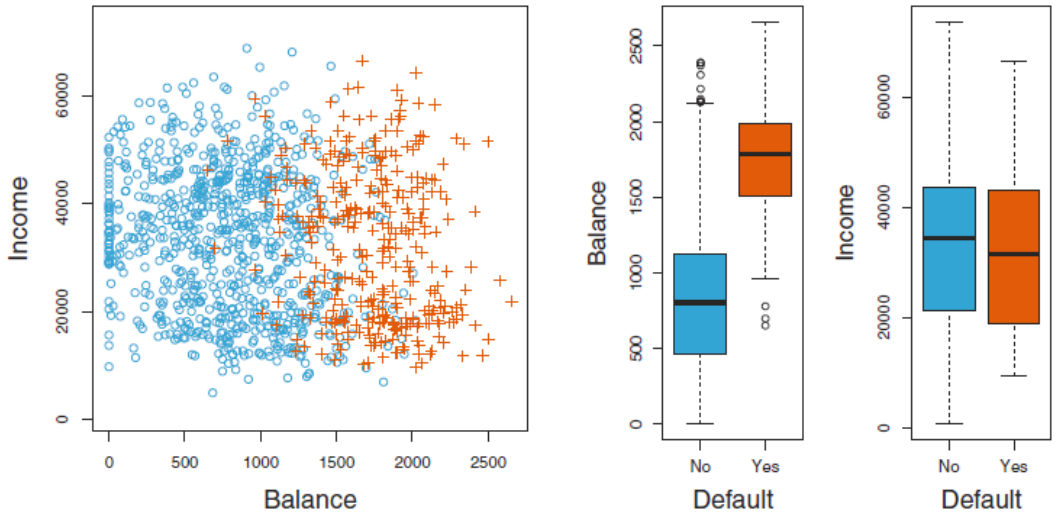
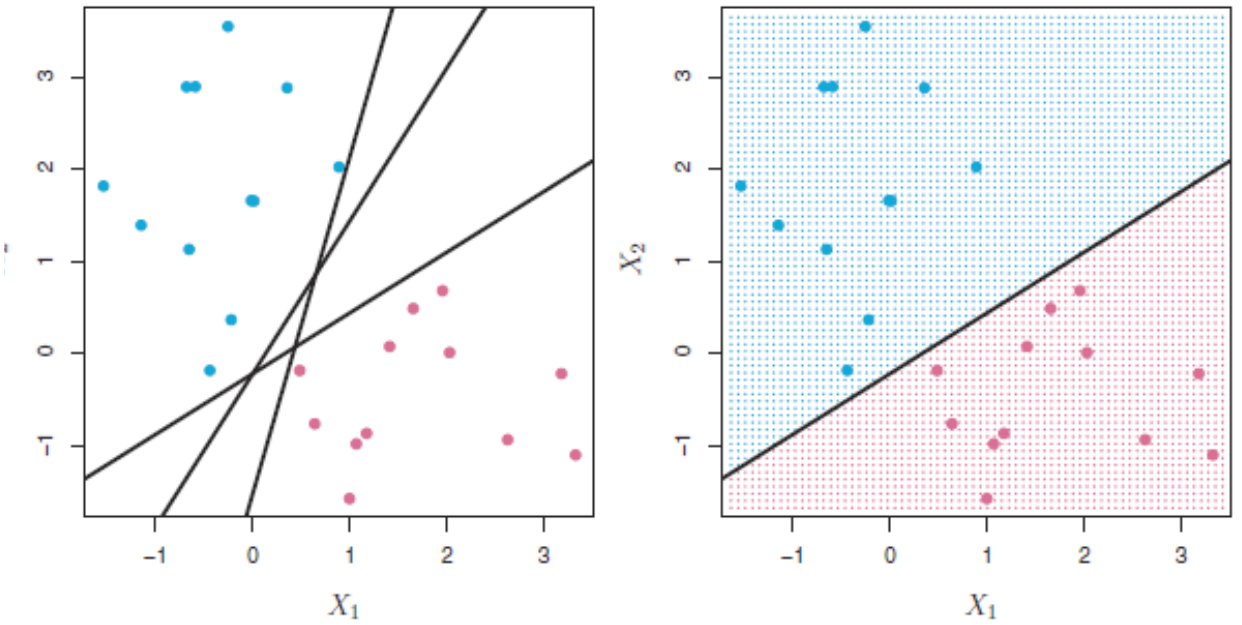


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

Predictor: Income and Balance (X)
Response: Default or not (Y)

Classification

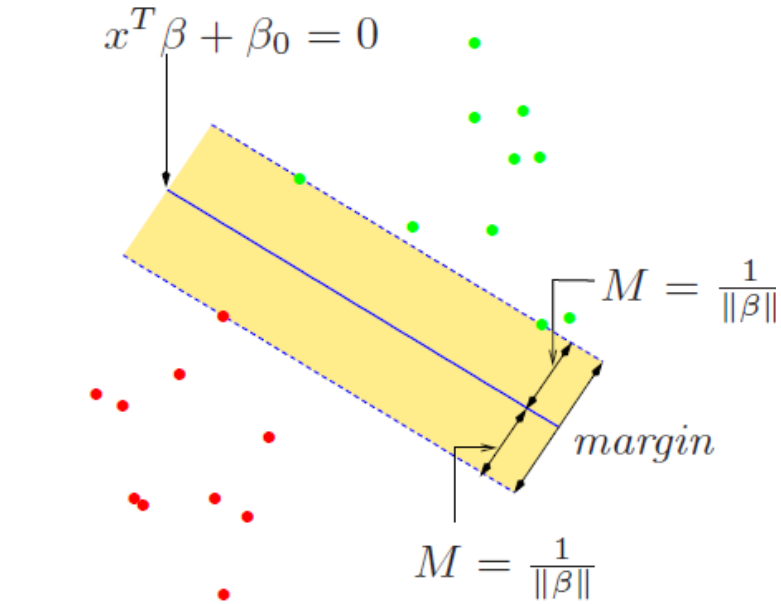
How to classify? Draw a hyperplane $X'\beta + \beta_0 = 0$



IGURE 9.2. Left: *There are two classes of observations, shown in blue and purple, each of which has measurements on two variables. Three separating perplanes, out of many possible, are shown in black.* Right: *A separating hyperplane is shown in black. The blue and purple grid indicates the decision rule ade by a classifier based on this separating hyperplane: a test observation that lls in the blue portion of the grid will be assigned to the blue class, and a test servation that falls into the purple portion of the grid will be assigned to the rple class.*

Support Vector Machine

How to find the optimal hyperplane $X'\beta + \beta_0 = 0$? when perfectly separable



Objective: Maximize margin = Minimize $||\beta||$
Constraints: separate the data correctly

Support Vector Machine

If $X_i'\beta + \beta_0 \geq 1,\ y_i = 1$ If $X_i'\beta + \beta_0 \leq -1,\ y_i = -1$

So the constraint is

$$y_i(X_i'\beta + \beta_0) \geq 1 \quad \forall i = 1, \ldots, n$$

The optimization problem is

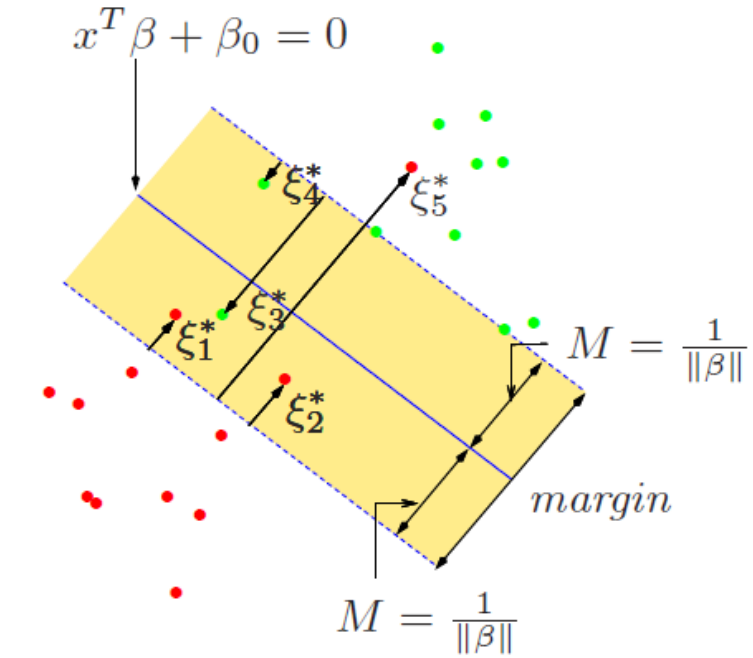
$$\min_{\beta, \beta_0} ||\beta||$$
$$s.t.\ y_i(X_i'\beta + \beta_0) \geq 1 \quad \forall i = 1, \ldots, n$$

or

$$\min_{\beta, \beta_0} ||\beta||^2$$
$$s.t.\ y_i(X_i'\beta + \beta_0) \geq 1 \quad \forall i = 1, \ldots, n$$

Support Vector Machine

When not perfectly separable



Objective: Minimize $||\beta||$ and errors
Constraints: separate the data correctly with possible errors

Support Vector Machine

When not perfectly separable, the optimization problem is

$$\min_{\beta,\beta_0} ||\beta||^2 + \lambda \sum_{i=1}^{n} \xi_i$$
$$s.t. \quad y_i(X_i'\beta + \beta_0) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n$$
$$\xi_i \geq 0 \quad \forall i = 1, \ldots, n$$

How to find the dual function? Recall the Lagrangian dual and KKT conditions.

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.
- Bruce E. Hansen, Lecture Notes on Nonparametrics `http://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf`