Linear Regression and *k*-NN
(Part A)

Contents

- Univariate Linear Regression
- Multivariate Linear Regression
- Nearest Neighbor Models
- Implementation and Experiments

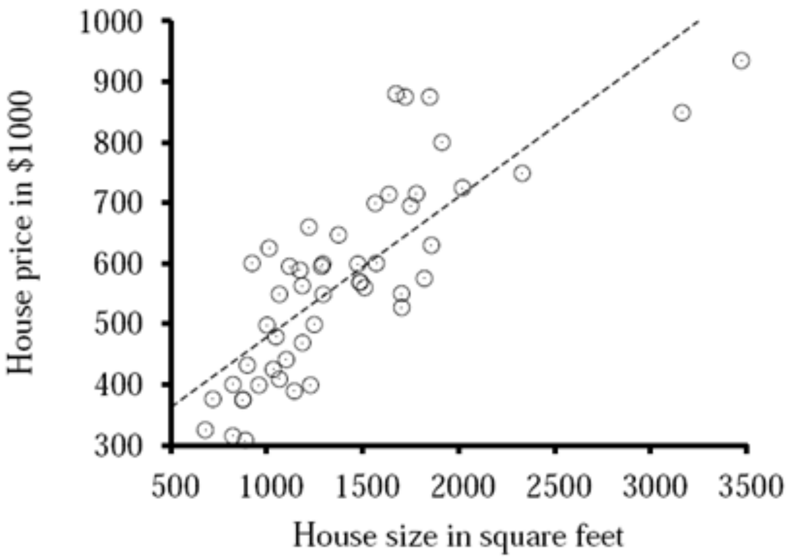
Univariate Linear Regression

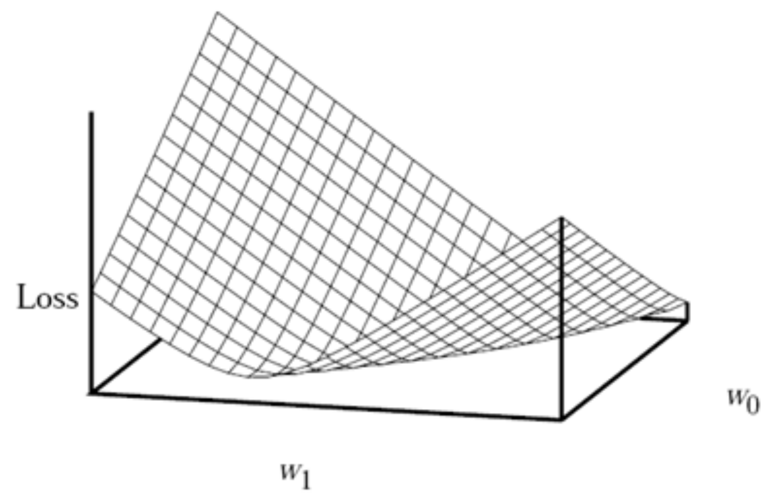
- The task of finding $h_{\mathbf{w}}$ that best fits the data

  $$h_{\mathbf{w}}(x) = w_1 x + w_0$$

- Need to find the values of the weights $[w_0, w_1]$ (= $\mathbf{w}$) that minimize the sum of squared error over all the training examples:

$$\sum_{j=1}^{N} (y_j - h_{\mathbf{w}}(x_j))^2 = \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2$$





Univariate Linear Regression

- To minimize the sum $\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))^2$

$$\frac{\partial}{\partial w_0}\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))^2=-2\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))=0$$

$$\frac{\partial}{\partial w_1}\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))^2=-2\sum_{j=1}^{N}(y_j-(w_1x_j+w_0))x_j=0$$

These equations have a unique solution:

$$w_0=\left(\sum y_j-w_1\left(\sum x_j\right)\right)/N; \qquad w_1=\frac{N\left(\sum x_jy_j\right)-\left(\sum x_j\right)\left(\sum y_j\right)}{N\left(\sum x_j^2\right)-\left(\sum x_j\right)^2}$$

Multivariate Linear Regression

- Each example $\mathbf{x}_j$ is an d-element vector

$$h_{\mathbf{w}}(\mathbf{x}_j) = \mathbf{w} \cdot \mathbf{x}_j = \mathbf{w}^T \mathbf{x}_j = \sum_i w_i x_{j,i}$$

  where $x_{j,0} = 1$ is a dummy input attribute

- The best weight vector minimizes loss over the examples:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j (y_j - \mathbf{w} \cdot \mathbf{x}_j)^2$$

- The analytical solution is

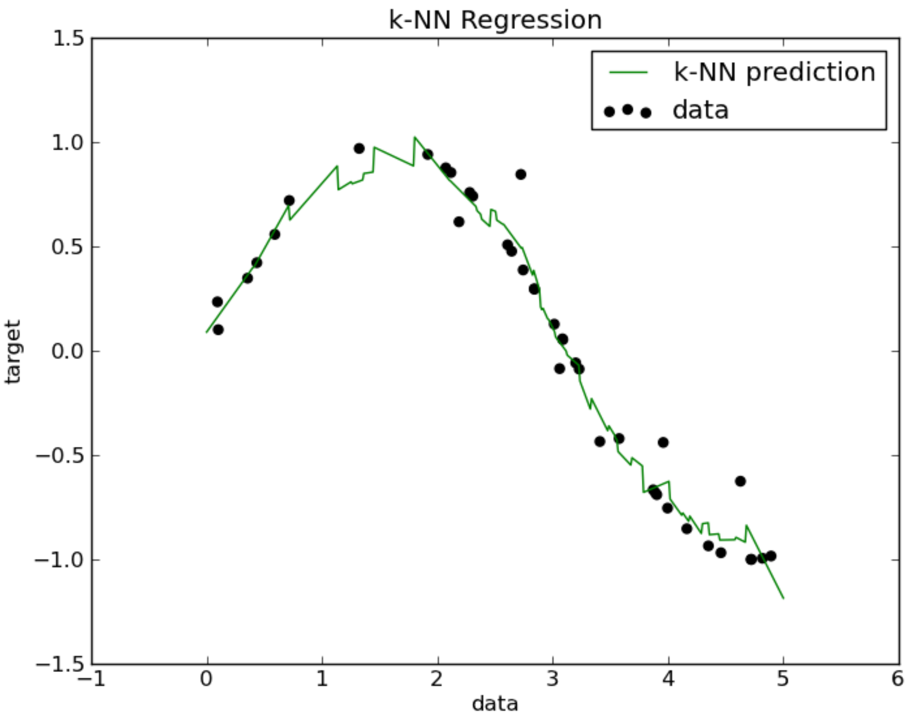
$$\mathbf{w}^* = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

  where $\mathbf{X}$ is the data matrix of inputs with one d-dimensional example per row

  – jth row contains $d + 1$ feature values including $x_{j,0} = 1$

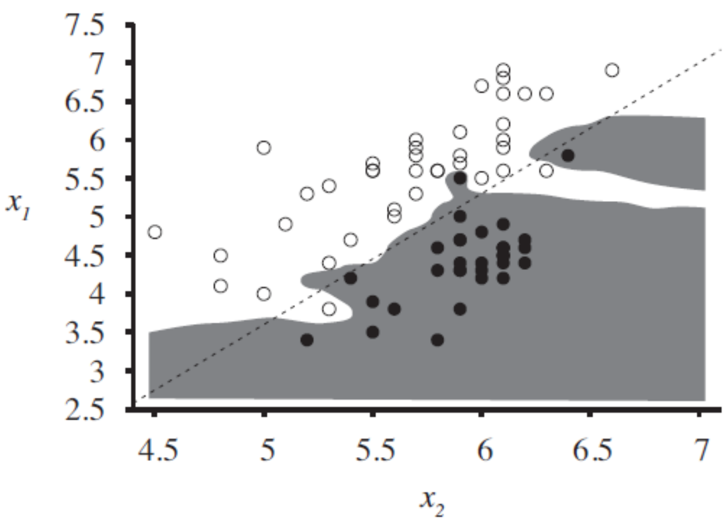
Nearest Neighbor Models

- Given a query $\mathbf{x}_q$, find the k nearest neighbors $NN(k, \mathbf{x}_q)$
  - Regression: mean or median of $NN(k, \mathbf{x}_q)$ or
    solve a linear regression problem on $NN(k, \mathbf{x}_q)$

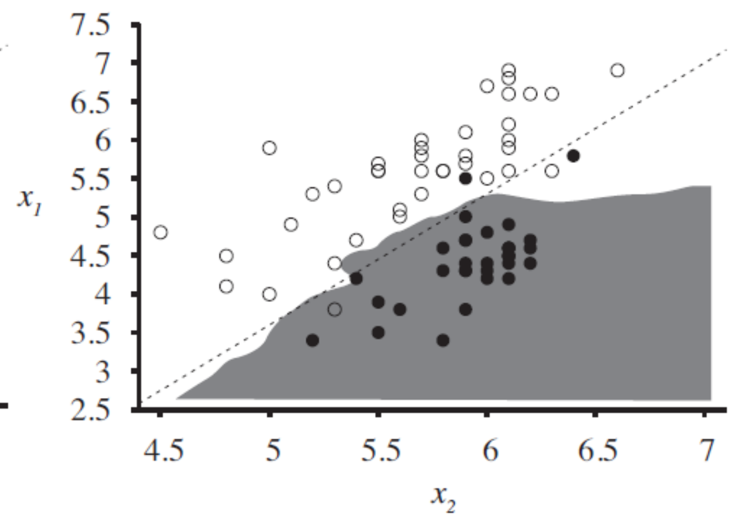


Nearest Neighbor Models

- Given a query $\mathbf{x}_q$, find the k nearest neighbors $NN(k, \mathbf{x}_q)$
  - Classification: plurality vote of $NN(k, \mathbf{x}_q)$



Overfitting with $k = 1$



O.K. with $k = 5$

Nearest Neighbor Models

- Distance metric: Minkowski distance (L^p norm)

$$L^p(\mathbf{x}_j, \mathbf{x}_q) = \left(\sum_i \left| x_{j,i} - x_{q,i} \right|^p \right)^{1/p}$$

  - $p = 2$: Euclidean distance
  - $p = 1$: Manhattan distance
  - $p = 1$ with Boolean attributes: Hamming distance

- Normalization:

$$x_{j,i} \to (x_{j,i} - \mu_i)/\sigma_i$$

  - μ_i: mean of the values in the ith dimension
  - σ_i: standard deviation of the values in the ith dimension