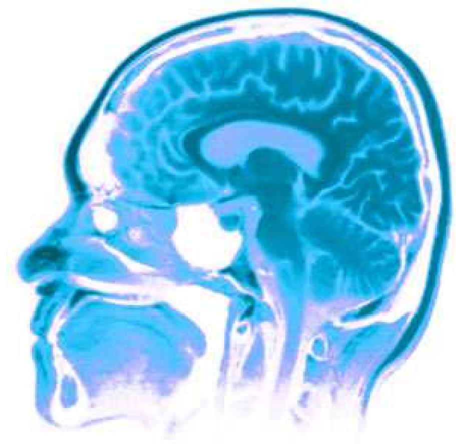
CPSC340



Linear prediction

Nando de Freitas
October, 2012
University of British Columbia

Outline of the lecture

This lecture introduces us to the topic of **supervised learning**. Here the data consists of **input**-**output** pairs. Inputs are also often referred to as **covariates**, **predictors** and **features**; while outputs are known as **variates** and **labels**. The goal of the lecture is for you to:

- ❑ Understand the supervised learning setting.
- ❑ Understand linear regression (aka **least squares**)
- ❑ Understand how to apply linear regression models to make predictions.
- ❑ Learn to derive the least squares estimate by optimization.

Linear supervised learning

- Many real processes can be approximated with linear models.
- Linear regression often appears as a module of larger systems.
- Linear problems can be solved analytically.
- Linear prediction provides an introduction to many of the core concepts of machine learning.

We are given a training dataset of n instances of input-ouput pairs $\{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$. Each input $\mathbf{x}_i \in \mathbb{R}^{1\times d}$ is a vector with d attributes. The inputs are also known as predictors or covariates. The output, often referred to as the target, will be assumed to be univariate, $\mathbf{y}_i \in \mathbb{R}$, for now.

A typical dataset with $n = 4$ instances and 2 attributes would look like the following table:

Wind speed	**People inside building**	**Energy requirement**
100	2	5 ← th. settings
50	42	25
45	31	22
60	35	18

day 1 $x_1 = (100, 2)$ $y_1 = 5$

$\boxed{y(x)}$

Energy demand prediction

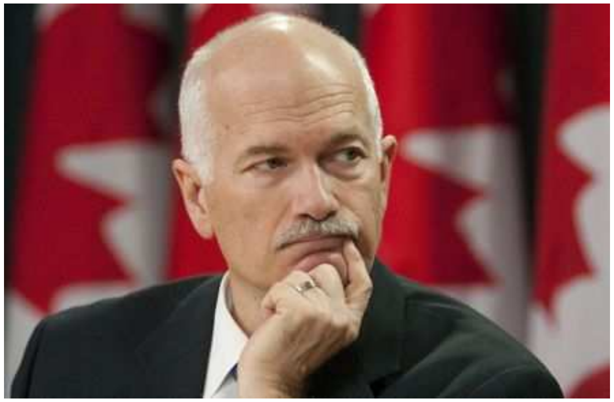
Given the training set $\{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$, we would like to learn a model of how the inputs affect the outputs. Given this model and a new value of the input $\mathbf{x}_{n+1}$, we can use the model to make a prediction $\widehat{y}(\mathbf{x}_{n+1})$.

1. Training: $\boxed{x_{1:n}, y_{1:n}} \rightarrow \boxed{\text{Learning mach.}} \rightarrow \hat{\Theta}$

2. Validation prediction: x_{n+1}, x_{n+2}, $\hat{\Theta} \rightarrow \boxed{\text{Predict machine}} \rightarrow \hat{y}_{n+1} \quad \hat{y}_{n+2}$

Prostate cancer example

❑ Goal: Predict a prostate-specific antigen (log of lpsa) from a number of clinical measures in men who are about to receive a radical prostatectomy.



❑The inputs are:

- Log cancer volume (lcavol)
- Log prostate weight (lweight)
- Age
- Log of the amount of benign prostatic hyperplasia (lbph)
- Seminal vesicle invasion (svi) - *binary*
- Log of capsular penetration (lcp)
- Gleason score (gleason) – *ordered categorical*
- Percent of Gleason scores 4 or 5 (pgg45)

Which inputs are more important?

website

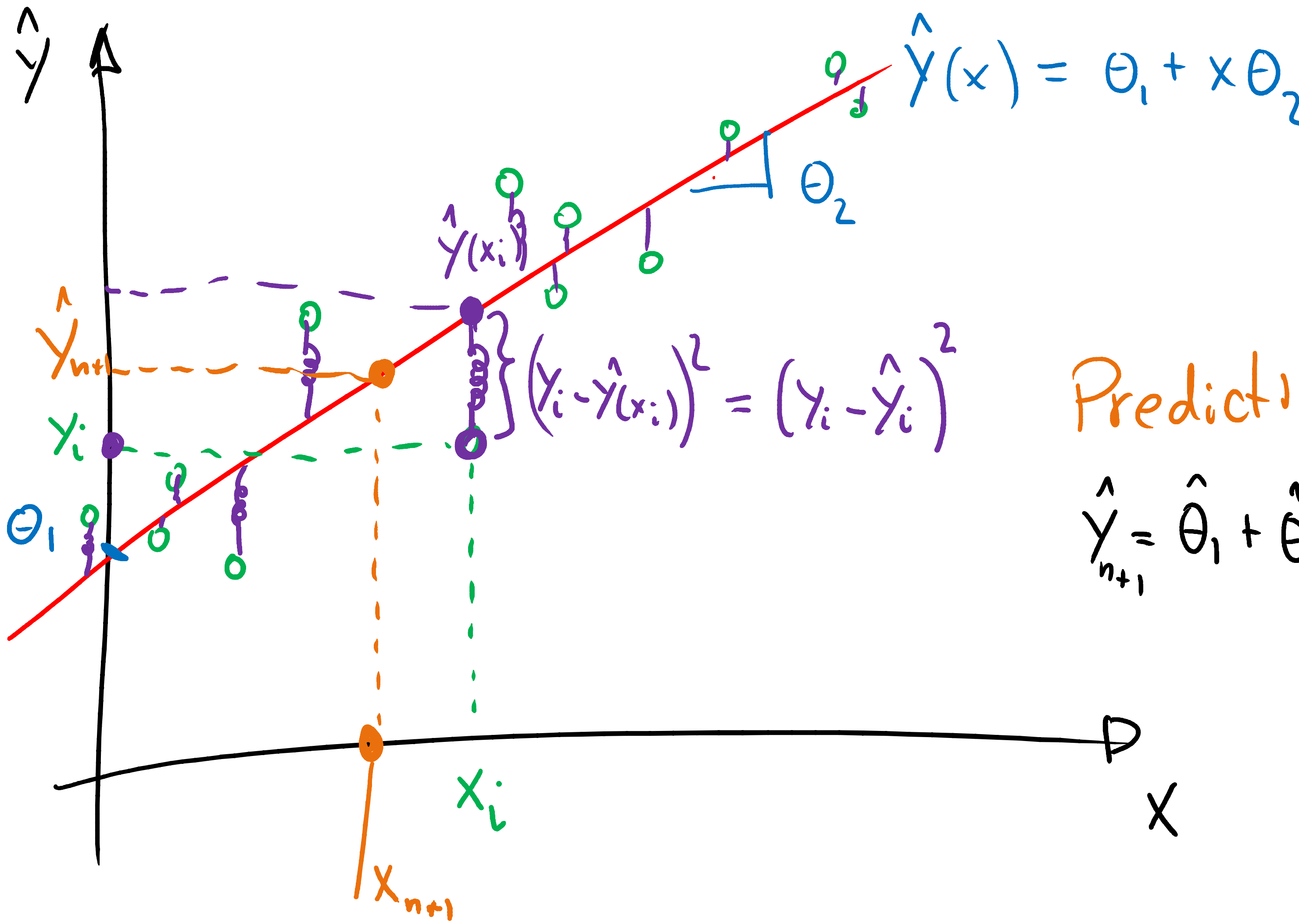
[Hastie, Tibshirani & Friedman book]

$$\hat{y}(\mathbf{x}_i) = \theta_1 + x_i\theta_2$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_1 - x_i\theta_2)^2$$

Goal: solve for θ_1 and θ_2

$\boxed{d=1}$



Prediction:

$$\hat{y}_{n+1} = \hat{\theta}_1 + \hat{\theta}_2 x_{n+1}$$

x_{i2} = heigh
x_{i3} = weight

Linear prediction

$y = \theta_0 + \theta_1 x_2$

$y = \sum_{j=1} \theta_j x_j, \quad x_1 = 1$

In general, the linear model is expressed as follows:

$Y = 1\theta_1 + x_{i2}\theta_2 + x_{i3}\theta_3$

$$\boxed{\widehat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j} = x_{i1}\theta_1 + x_{i2}\theta_2 + \cdots x_{id}\theta_d$$

where we have assumed that $x_{i1} = 1$ so that θ_1 corresponds to the intercept of the line with the vertical axis. θ_1 is known as the bias or offset.

In matrix form, the expression for the linear model is:

$$\boxed{\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta},}$$

with $\widehat{\mathbf{y}} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$. That is,

$$\begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}.$$

Wind speed	People inside building	Energy requirement
100	2	5
50	42	25
45	31	22
60	35	18

For our energy prediction example, we would form the following matrices with $n = 4$ and $d = 3$:

$$\mathbf{y} = \begin{bmatrix} 5 \\ 25 \\ 22 \\ 18 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}.$$

Suppose that $\hat{\boldsymbol{\theta}} = [1 \quad 0 \quad 0.5]^T$. Then, by multiplying $\mathbf{X}$ times $\hat{\boldsymbol{\theta}}$, we would get the following predictions on the training set:

$$\widehat{\mathbf{y}} = \begin{bmatrix} 2 \\ 22 \\ 16.5 \\ 18.5 \end{bmatrix} = \overset{X}{\begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix}} \overset{\hat{\theta}}{\begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}}.$$

Linear prediction

Likewise, for a point that we have never seen before, say x = [50 20], we generate the following prediction:

$$\hat{y}(x_{n+1}) = \underbrace{[1\ 50\ 20]}_{X_{n+1}} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}}_{\hat{\theta}} = 1 + 0 + 10 = 11.$$

Optimization approach

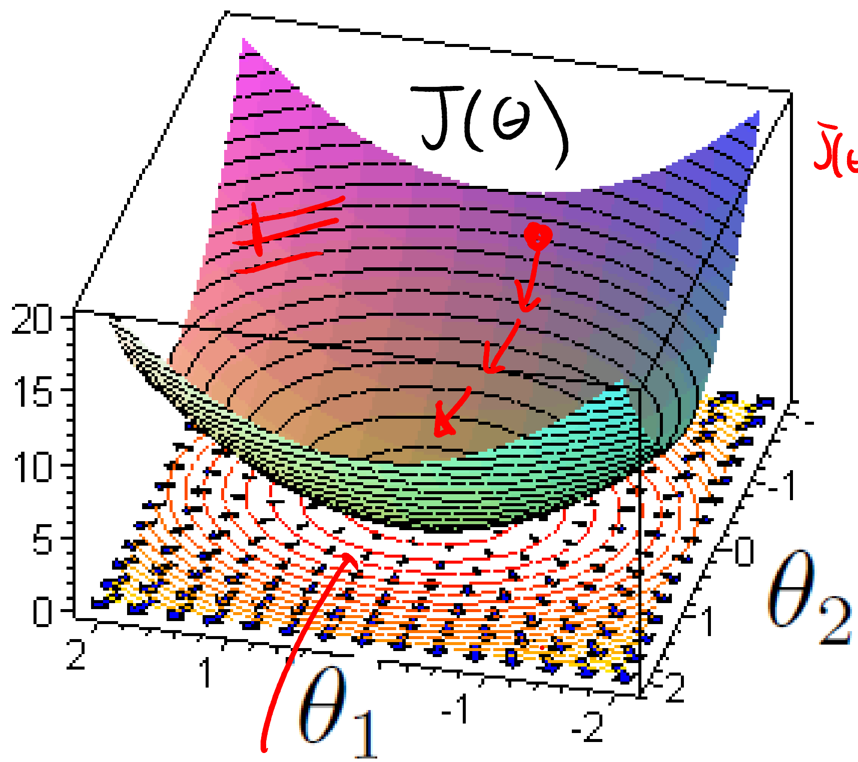
$\hat{y}_1 = x_1^T \theta$

Our aim is to mininimise the quadratic cost between the output labels and the model predictions $\|Y - X\theta\|^2$

$x_1^T = (x_{11} \;\; x_{12})$

$x_2^T = (x_{21} \;\; x_{22})$

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2$$



Contours

$$J(\theta) = \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right)^T \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right)$$

$$= \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} x_1^T\theta \\ x_2^T\theta \end{bmatrix}\right)^T \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{matrix} x_1^T\theta \\ x_2^T\theta_2 \end{matrix}\right)$$

$$= \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}\right)^T \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}\right)$$

$$= \left[(y_1 - \hat{y}_1) \;\; (y_2 - \hat{y}_2)\right] \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \end{bmatrix} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2$$

Optimization

$$J(\boldsymbol{\theta}) = \underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T}_{1\times n}\underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})}_{n\times 1} = \underbrace{\sum_{i=1}^{n}\left(y_i - x_i^T\theta\right)^2}_{1\times 1}$$

We will need the following results from matrix differentiation:

$\boxed{\frac{\partial \mathbf{A}\theta}{\partial \boldsymbol{\theta}} = \mathbf{A}^T}$ and $\frac{\partial \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{A}^T\boldsymbol{\theta}$

$A = y^T X$

$A = X^T X$

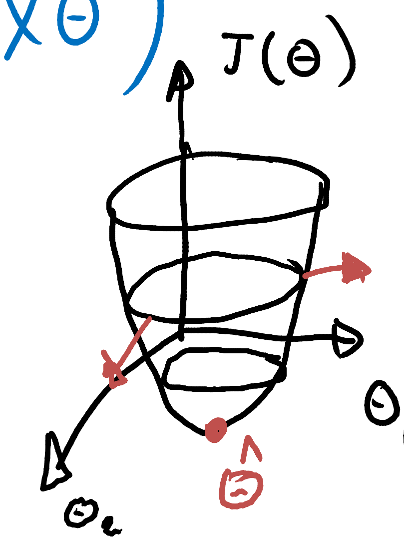
$(X\theta)^T = \theta^T X^T$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial\theta}\left((y - X\theta)^T(y - X\theta)\right)$$

$$\overset{(2)}{=} \frac{\partial}{\partial\theta}\left(y^Ty \underbrace{- \overset{1\times 1}{y^TX\theta} - \overset{1\times 1}{\theta^T X^T y}}_{-2y^TX\theta} + \theta^T X^T X\theta\right)$$

$$\overset{(3)}{=} 0 - 2\left(y^TX\right)^T + 2\left(X^TX\right)^T\theta$$

$$= -2X^Ty + 2X^TX\theta = \nabla J(\theta)$$



Least squares estimates

$$-2X^Ty + 2X^TX\theta = 0$$

$[X]^{-1}$ NO

$$\overset{d\times n}{X^T}\ \overset{n\times d}{X}\ \overset{d\times 1}{\theta} = \overset{d\times n}{X^T}\ \overset{n\times 1}{y}$$

$\frac{2}{X}$ NO

$$\boxed{\hat{\theta}_{LS} = (X^TX)^{-1}X^Ty}$$

$$\boxed{\hat{y} = \underbrace{X(X^TX)^{-1}X^T}_{\text{Hat}}y}$$

$$\hat{y} = x^*\hat{\theta}_{LS}$$

$$X^TX\theta = X^Ty$$

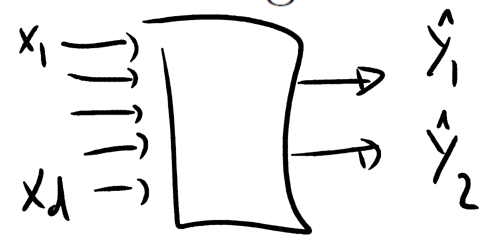
$$\underbrace{(X^TX)^{-1}X^TX}_{I}\theta = (X^TX)^{-1}X^Ty$$

$$I\theta = (X^TX)^{-1}X^Ty$$

Multiple outputs

If we have several outputs $\mathbf{y}_i \in \mathbb{R}^c$, our linear regression expression becomes:

eg $c=2$



$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} \\ \vdots & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1d} \\ 1 & & & \\ 1 & & \ddots & \\ \vdots & & & \\ 1 & & & x_{nd} \end{bmatrix} \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \vdots & \vdots \\ \Theta_{d1} & \Theta_{d2} \end{bmatrix}$$

Next lecture

In the next lecture, we learn to derive the linear regression estimates by maximum likelihood with multivariate Gaussian distributions.