Linear Regression

Maria-Florina Balcan

03/25/2019

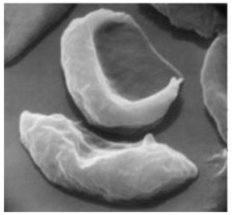
From Discrete to Continuous Labels

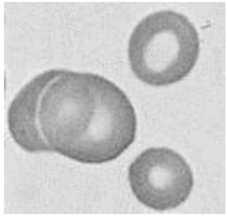
Classification

Sports
Science
News

X = Document

Y = Topic





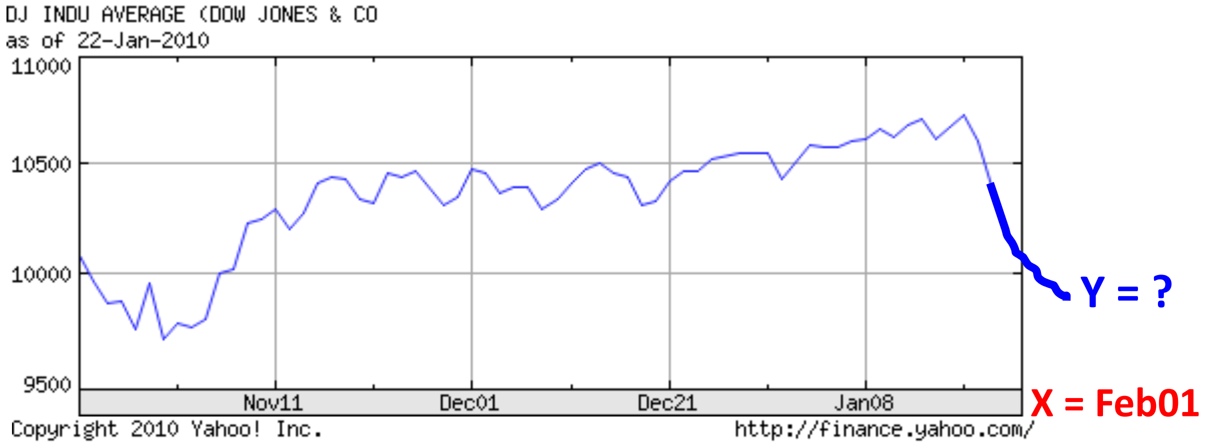
Anemic cell
Healthy cell

X = Cell Image

Y = Diagnosis

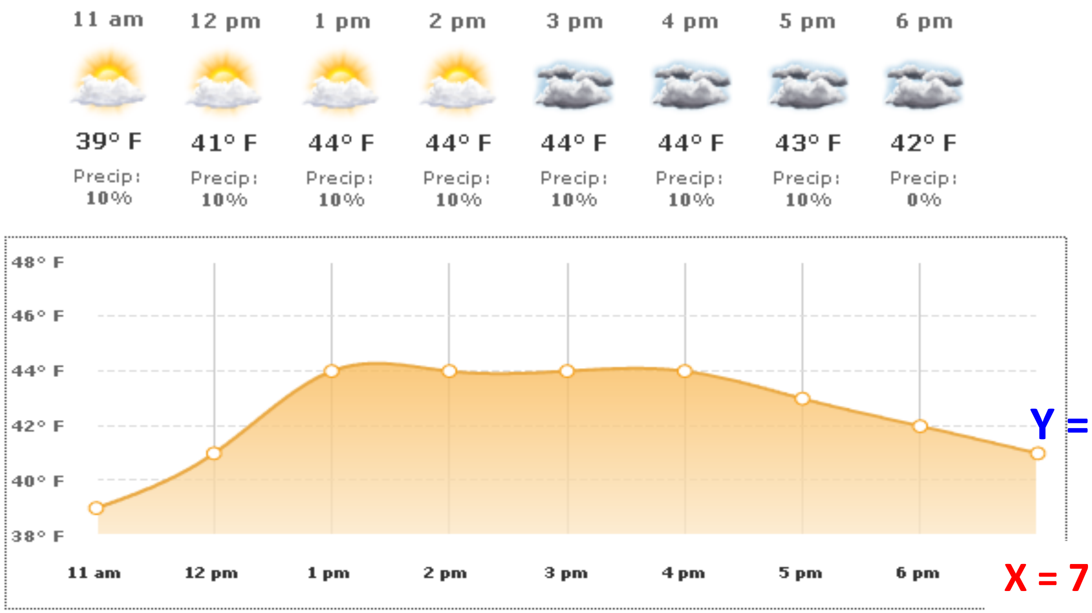
Regression

Stock Market
Prediction



Regression Tasks

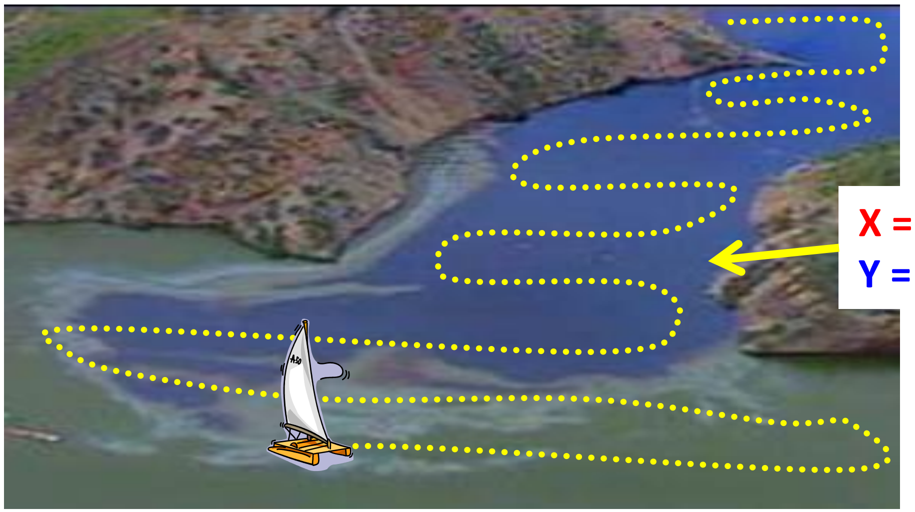
Weather Prediction



Y = Temp

X = 7 pm

Estimating Contamination



X = new location

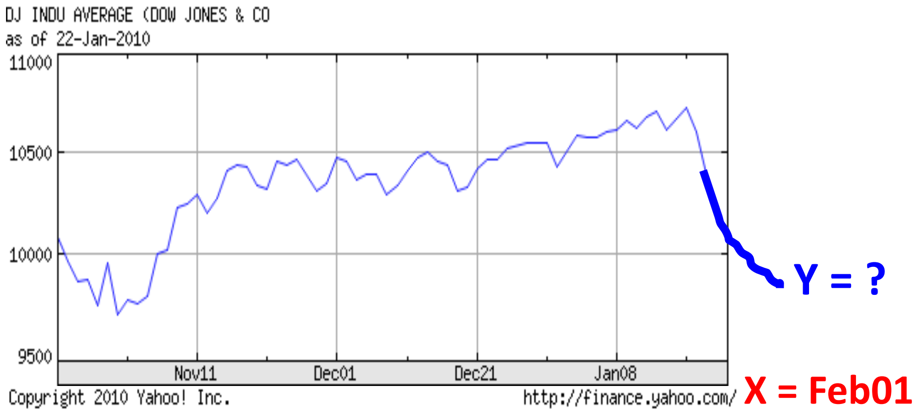
Y = sensor reading

Supervised Learning

Goal: Construct a **predictor** $f: X \rightarrow Y$ to minimize a risk (error measure) $err(f)$.

Typical Error Measures

Sports
Science
News



Classification:

$$err(f) = P(f(X) \neq Y)$$

Probability of Error

Regression:

$$err(f) = E[(f(X) - Y)^2]$$

Mean Squared Error

Linear Regression

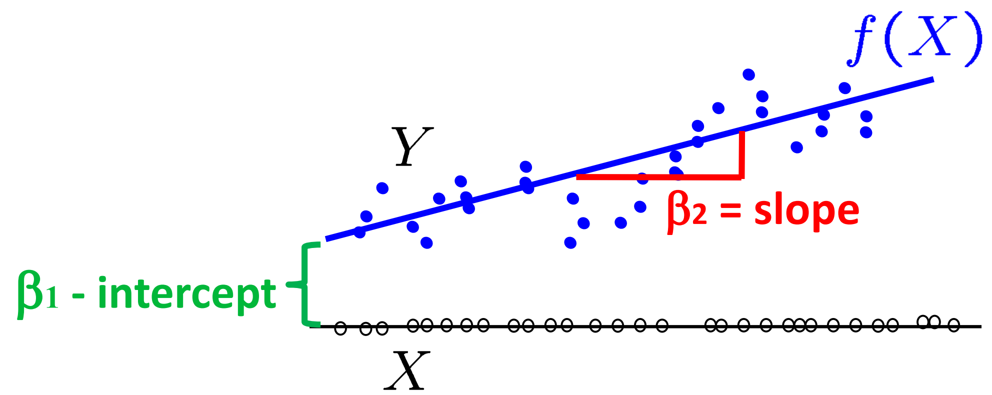
$$\hat{f}_n^L = \arg\min_{f \in F_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

Least Squares Estimator

F_L - Class of Linear functions

Uni-variate case:

$$f(X) = \beta_1 + \beta_2 X$$



Multi-variate case:

$$f(X) = f\left(X^{(1)}, \dots, X^{(p)}\right) = \beta_1 \cancelto{1}{X^{(1)}} + \beta_2 X^{(2)} + \cdots + \beta_p X^{(p)}$$

$$= X\beta \quad \text{where} \quad X = \left[X^{(1)} \dots X^{(p)}\right], \qquad \beta = \left[\beta_1 \dots \beta_p\right]^T$$

Least Squares Estimator

$$\hat{f}_n^L = \arg\min_{f \in F_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

$$\Downarrow$$

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (X_i \beta - Y_i)^2 \qquad\qquad \hat{f}_n^L(X) = X\hat{\beta}$$

$$= \arg\min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y})$$

$$\mathbf{A} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}$$

Least Squares Estimator

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^{T} (\mathbf{A}\beta - \mathbf{Y}) = \arg\min_{\beta} J(\beta)$$

$$J(\beta) = \quad (\mathbf{A}\beta - \mathbf{Y})^{T} (\mathbf{A}\beta - \mathbf{Y})$$

$$\left.\frac{\partial J(\beta)}{\partial \beta}\right|_{\hat{\beta}} = 0$$

Normal Equations

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\hat{\beta} = \mathbf{A}^{\mathrm{T}}\mathbf{Y}$$

p xp   p x1   p x1

If $(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ is invertible,

$$\boxed{\hat{\beta} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y} \qquad \hat{f}_n^{L}(X) = X\hat{\beta}}$$

Geometric Interpretation

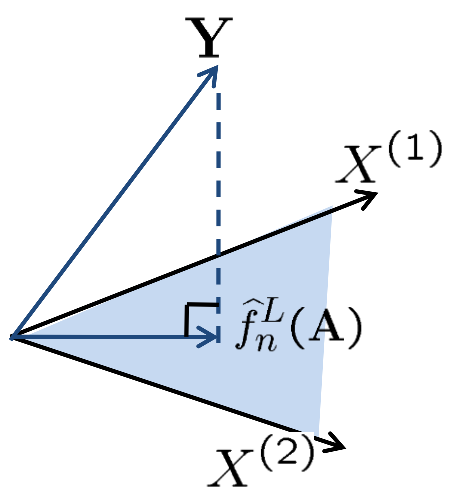
$$\hat{f}_n^L(X) = X\hat{\beta} = X\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Y}$$

Difference in prediction on training set:

$$\hat{f}_n^L(\mathbf{A}) - \mathbf{Y} =$$

$$\mathbf{A}^T\left(\hat{f}_n^L(\mathbf{A}) - \mathbf{Y}\right) = 0$$

$\hat{f}_n^L(\mathbf{A})$ is the orthogonal projection of $\mathbf{Y}$ onto the linear subspace spanned by the columns of $\mathbf{A}$.



Revisiting Gradient Descent

Even when $(\mathbf{A}^T\mathbf{A})$ is invertible, might be computationally expensive if $\mathbf{A}$ is huge.

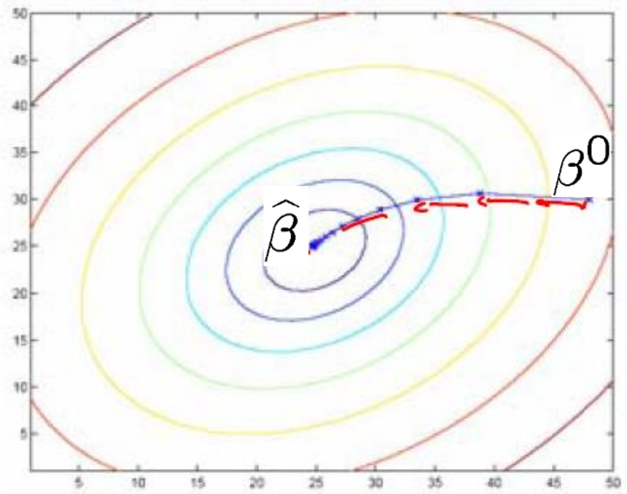
$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) = \arg\min_{\beta} J(\beta)$$

Gradient Descent since $J(\beta)$ is convex

Initialize: β^0

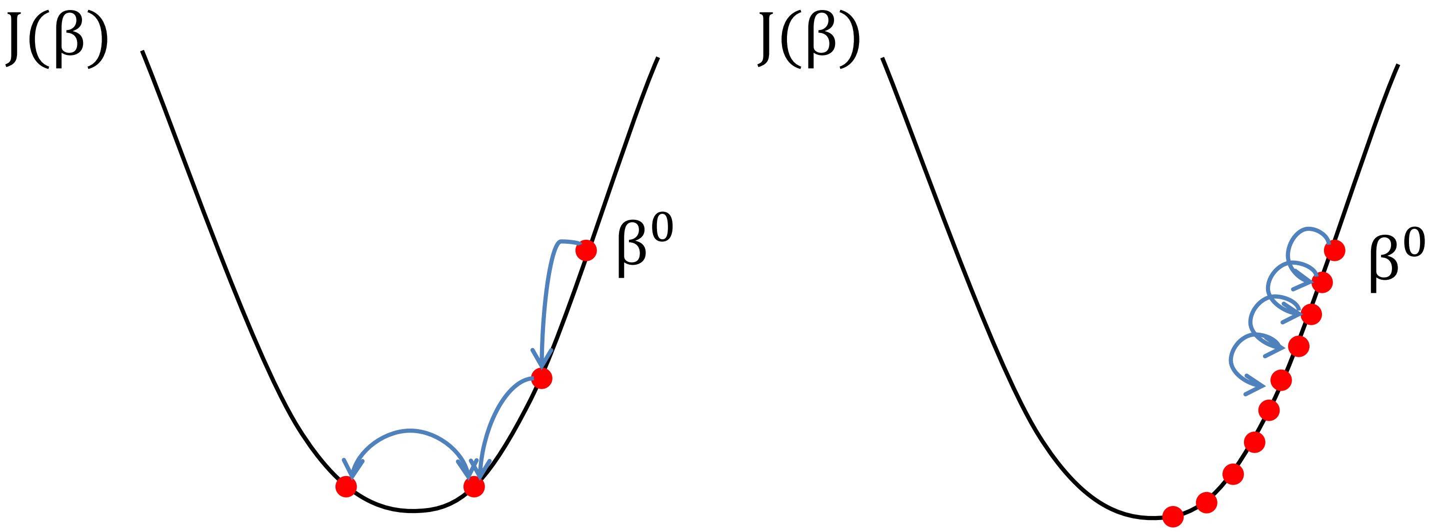
Update: $\beta^{t+1} = \beta^t - \frac{\alpha}{2}\frac{\partial J(\beta)}{\partial \beta}\Big|_t$

$$= \beta^t - \alpha \mathbf{A}^T \underbrace{(\mathbf{A}\beta^t - \mathbf{Y})}_{0 \text{ if } \beta^t = \hat{\beta}}$$



Stop: when some criterion met, e.g. fixed # iterations, or $\frac{\partial J(\beta)}{\partial \beta}\Big|_{\beta^t} < \epsilon$

Effect of step-size α



Large $\alpha \Rightarrow$ Fast convergence but larger residual error
Also possible oscillations

Small $\alpha \Rightarrow$ Slow convergence but small residual error

Least Squares and MLE

Intuition: Signal plus (zero-mean) Noise model

$$Y = f^*(X) + \epsilon = X\beta^* + \epsilon \qquad \epsilon \sim N(0, \sigma^2 \mathbf{I})$$

$$Y \sim N(X\beta^*, \sigma^2 \mathbf{I})$$

$$\hat{\beta}_{\mathrm{MLE}} = \arg\max_{\beta} \underbrace{\log p(\{(X_i, Y_i)\}_{i=1}^n \mid \beta, \sigma^2, X)}_{\text{log likelihood}}$$

$$= \arg\min_{\beta} \sum_{i=1}^{n} (X_i\beta - Y_i)^2 = \hat{\beta}$$

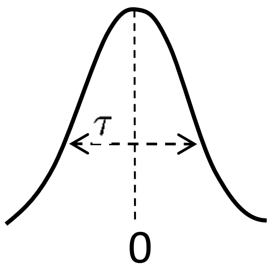
Least Square Estimate is same as Maximum Likelihood Estimate under a Gaussian model !

Regularized Least Squares and MAP

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \underbrace{\log p(\{(X_i, Y_i)\}_{i=1}^{n} \mid \beta\sigma^2)}_{\text{log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

I) Gaussian Prior

$$\beta \sim N(0, \tau^2 \mathbf{I}) \qquad p(\beta) \propto e^{-\beta^T\beta/2\tau^2}$$



$$\hat{\beta}_{MAP} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \underset{\substack{\downarrow \\ \text{constant}(\sigma^2, \tau^2)}}{\lambda} \|\beta\|_2^2$$

Ridge Regression

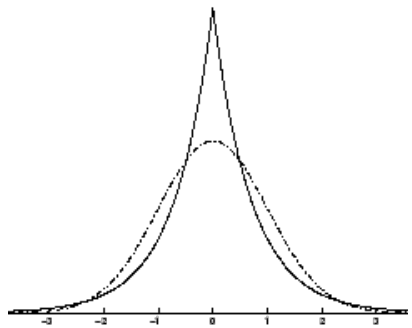
Prior belief that β is Gaussian with zero-mean biases solution to "small" β

Regularized Least Squares and MAP

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \underbrace{\log p(\{(X_i, Y_i)\}_{i=1}^{n} \mid \beta\sigma^2)}_{\text{log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

II) Laplace Prior

$\beta_i \sim \text{Laplace}(0, t)$ [iid] $\qquad p(\beta_i) \propto e^{-|\beta_i|/t}$



$$\hat{\beta}_{MAP} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \underset{\downarrow \atop \text{constant}(\sigma^2, t)}{\lambda} \|\beta\|_1$$

Lasso

Prior belief that β is Laplace with zero-mean biases solution to “small” β