Regression

Amr

1/28

Slides Credit: Aarti's Lecture slides and Eric's Lecture slides

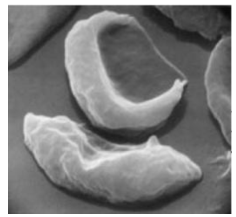
Big Picture

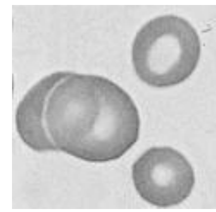
- Supervised Learning
  - Classification
    - Input x: feature vector
    - Output: discrete class label
  - Regression
    - Input x: feature vector
    - Output y: continuous value

Classification Tasks

	Features, X		Labels, Y

Diagnosing sickle cell anemia





Anemic cell
Healthy cell

Tax Fraud Detection

Refund	Marital Status	Taxable Income
No	Married	80K

Cheat
?

Web Classification

Sports
Science
News

Predict squirrel hill resident

Drive to CMU, Rachel's fan, Shop at SH Giant Eagle

Resident
Not resident

Classification

Goal: Construct a **predictor** $f : X \to Y$ to minimize a risk (performance measure) $R(f)$

Sports
Science
News

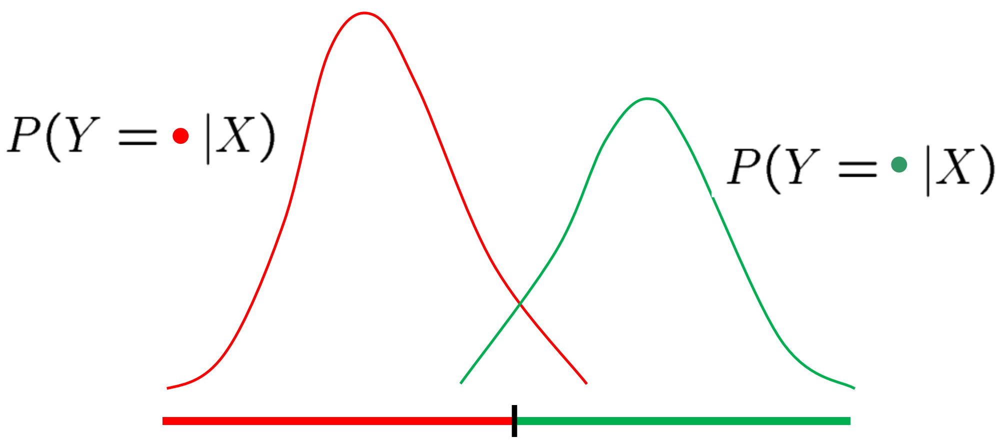
Features, X **Labels, Y**

$$R(f) = P(f(X) \neq Y)$$ **Probability of Error**

Classification

Optimal predictor:
(Bayes classifier)

$$f^* = \arg\min_f P(f(X) \neq Y)$$



$$f^*(X) = \begin{cases} {\color{red}\bullet} & P(Y = {\color{red}\bullet} | X) > P(Y = {\color{green}\bullet} | X) \\ {\color{green}\bullet} & \text{otherwise} \end{cases}$$

Depends on **unknown** distribution P_{XY}

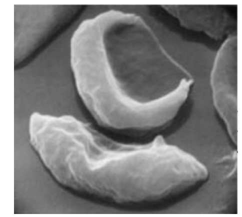
Discrete to Continuous Labels

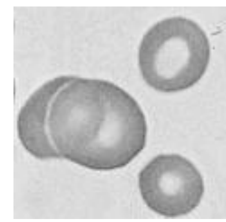
Classification

Sports
Science
News

X = Document

Y = Topic





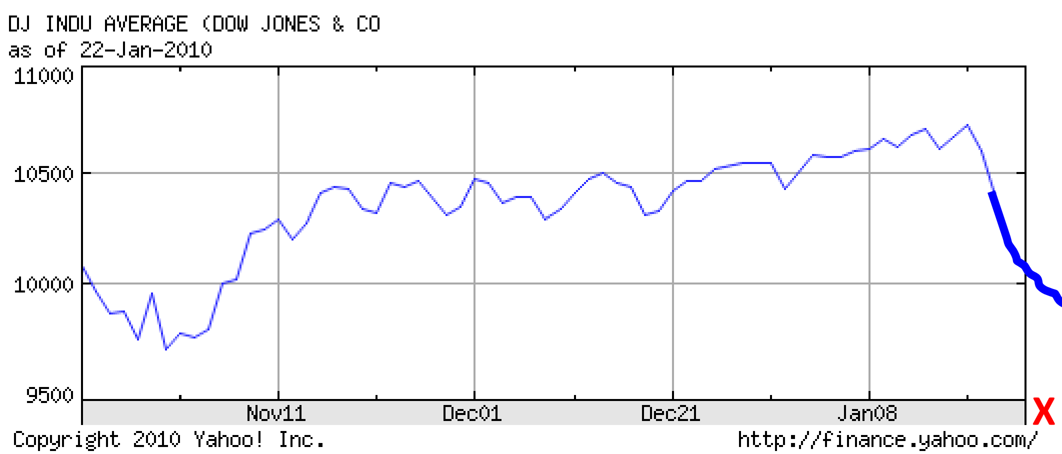
Anemic cell
Healthy cell

X = Cell Image

Y = Diagnosis

Regression

Stock Market Prediction



Y = ?

X = Feb01

Regression

- What is the equivalent of Bayes-optimal classifier?
- How about if we can model P(Y|X)?
- How can we predict Y given new X?
- We need a LOSS function
  - How about square loss?
  - What should be the prediction?

Regression (See board)

Optimal predictor:
(Conditional Mean)

$$f^* = \arg\min_f \mathbb{E}[(f(X) - Y)^2]$$

$$R(f) = \mathbb{E}_{XY}[(f(X) - Y)^2] = \mathbb{E}_X[\mathbb{E}_{Y|X}[(f(X) - Y)^2|X]]$$

Dropping subscripts for notational convenience

$$= E\left[E\left[(f(X) - E[Y|X] + E[Y|X] - Y)^2|X\right]\right]$$

$$= E[\quad E[(f(X) - E[Y|X])^2|X] \\ +2E\left[(f(X) - E[Y|X])(E[Y|X] - Y)|X\right] \\ +E[(E[Y|X] - Y)^2|X]]$$

$$= E[\quad E[(f(X) - E[Y|X])^2|X] \\ +2(f(X) - E[Y|X]) \times 0 \\ +E[(E[Y|X] - Y)^2|X]]$$

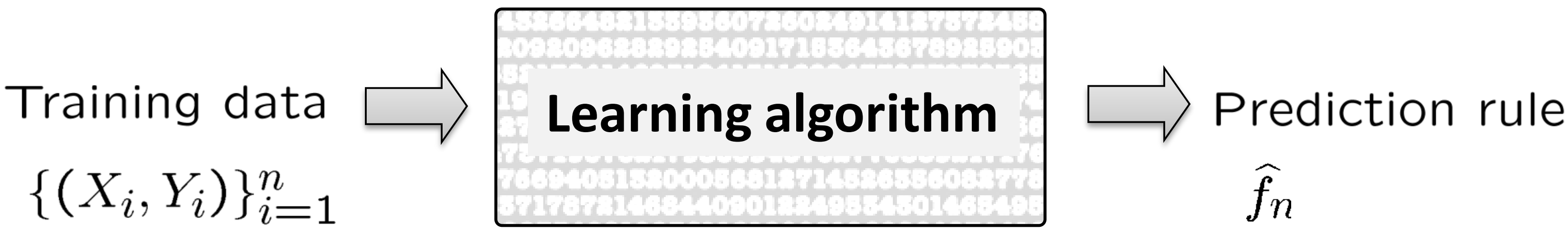
$$= E\left[(f(X) - E[Y|X])^2\right] + R(f^*).$$

Thus $R(f) \geq R(f^*)$ for any prediction rule f, and therefore $R^* = R(f^*)$.

Models

- So how can we proceed?
- We need to make some assumption to model P(Y|X)
  - Linear form (basis function)
  - Noise distribution
  - Loss function
  - Etc.

Regression algorithms



Linear Regression

Lasso, Ridge regression (Regularized Linear Regression)

Nonlinear Regression

Kernel Regression

Regression Trees, Splines, Wavelet estimators, …

Empirical Risk Minimizer:

$$\widehat{f}_n = \arg\min_f \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{Empirical mean}} (f(X_i) - Y_i)^2$$

Least Squares Estimator (on board)

$$\widehat{f}_n^L = \arg\min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

$$\Downarrow$$

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

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

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

Vector Derivative (see notes from website)

- Some useful facts: assume that A is symmetric

$$\nabla_x \left(a^T x \right) = a$$

$$\nabla_x \left(x^T a \right) = a$$

$$\nabla_x \left(Ax \right) = A^T$$

$$\nabla_x \left(x^T A x \right) = 2Ax$$

$$\nabla_x \left((A-x)^T A (A-x) \right) = -2A(A-x)$$

$$\nabla_x \left(x^T x \right) = 2x$$

Probabilistic Interpretation: MLE

Intuition: Signal plus (zero-mean) Noise model

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

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

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

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

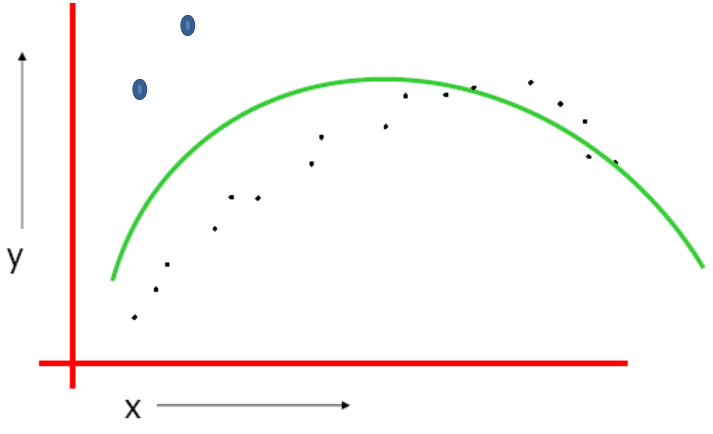
Least Square Estimate is same as Maximum Likelihood Estimate under a Gaussian model !

Variations

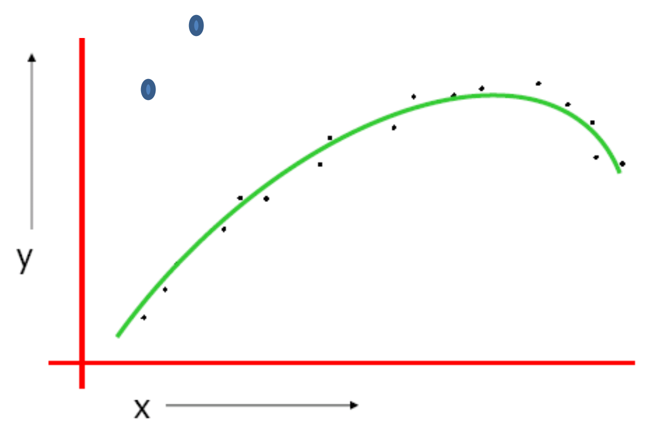
- What if the noise terms are independent but not identical?
  - Homework
- What if they are IID but not Gaussian?
- Think about robustness
  - What if we have outliers?

Robustness

- The best fit from a quadratic regression



- But this is probably better …



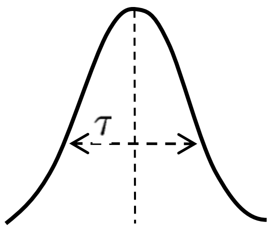
Regularized Least Squares and MAP

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

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

I) Gaussian Prior

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



$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \|\beta\|_2^2$$

Ridge Regression

$\lambda \rightarrow$ constant(σ^2, τ^2)

Closed form: HW

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

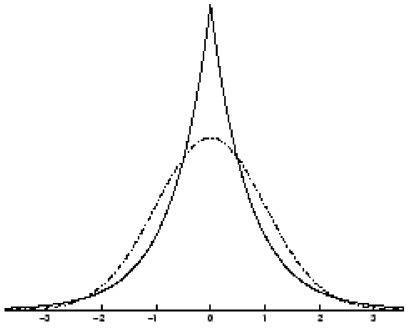
Regularized Least Squares and MAP

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

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

II) Laplace Prior

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



$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \|\beta\|_1$$

Lasso

$\lambda \downarrow$ constant(σ^2, t)

Closed form: HW

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

Ridge Regression vs Lasso

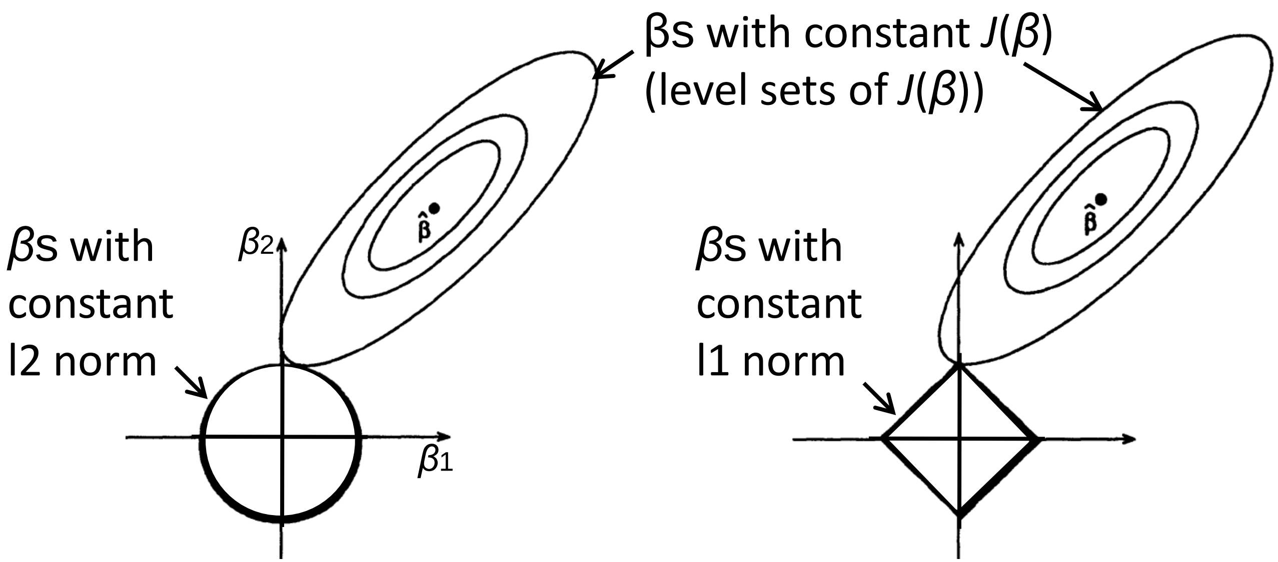
$$\min_{\beta}(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda \text{pen}(\beta) = \min_{\beta} J(\beta) + \lambda \text{pen}(\beta)$$

Ridge Regression:

$$\text{pen}(\beta) = \|\beta\|_2^2$$

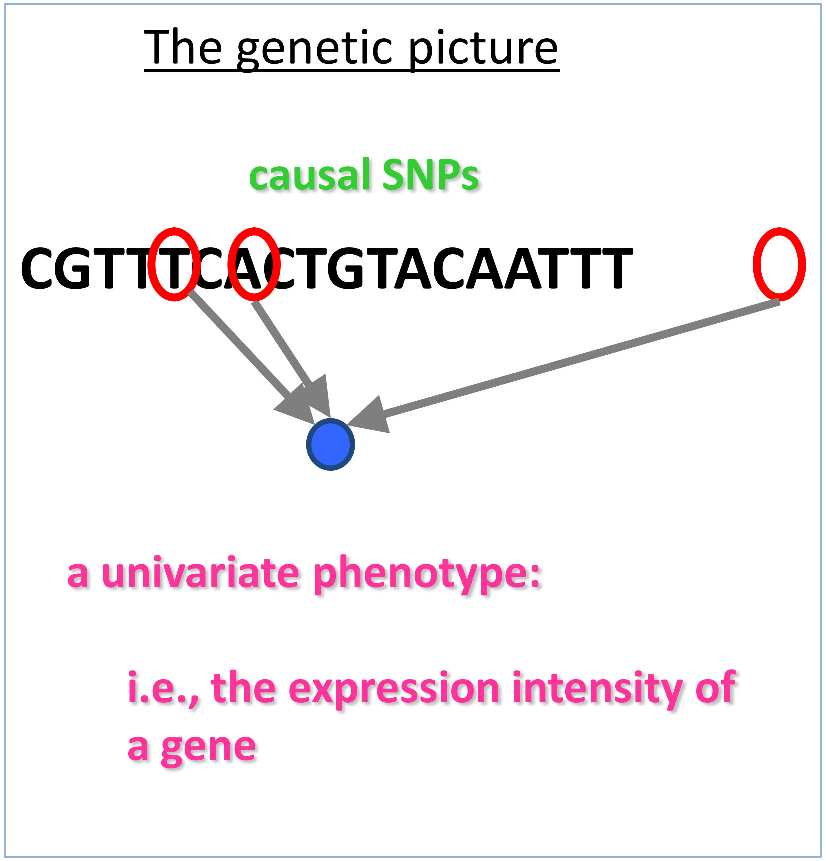
Lasso:

$$\text{pen}(\beta) = \|\beta\|_1$$



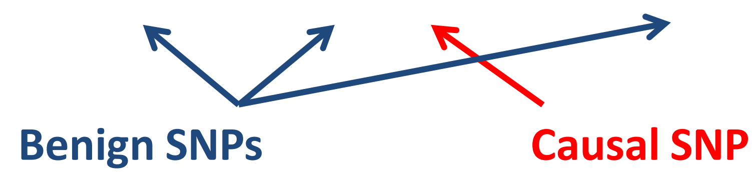
Lasso (l1 penalty) results in sparse solutions – vector with more zero coordinates
Good for high-dimensional problems – don't have to store all coordinates!

Case study:
predicting gene expression



Association Mapping as Regression

	Phenotype (BMI)	Genotype
Individual 1	2.5	. . C . . . . . T . . C . . . . T . . . . . C . . . . . A . . C . . . . T . . .
Individual 2	4.8	. . G . . . . . A . . G . . . . A . . . . . C . . . . . T . . C . . . . T . . .
⋮		
Individual N	4.7	. . G . . . . . T . . C . . . . T . . . . . G . . . . . T . . G . . . . T . . .



Association Mapping as Regression

	Phenotype (BMI)	Genotype
Individual 1	2.5	. . 0 . . . . . 1 . . 0 . . . . 0 . . .
Individual 2	4.8	. . 1 . . . . . . 1 . . 1 . . . . 1 . . .
⋮		
Individual N	4.7	. . 2 . . . . . . 2 . . 1 . . . . 0 . . .

$$\mathbf{y}_i = \sum_{j=1}^{J} x_{ij}\beta_j$$

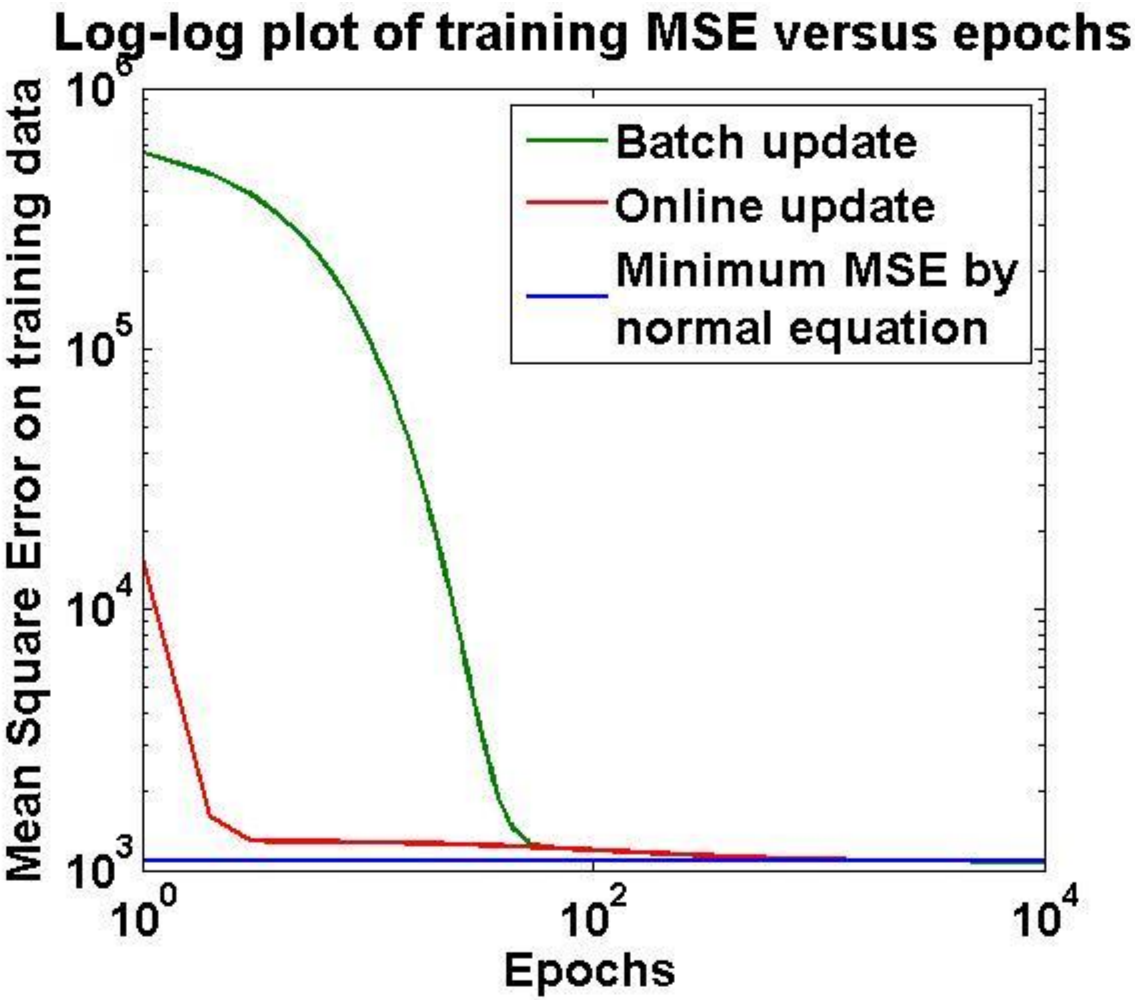
SNPs with large $|\beta_j|$ are relevant

Experimental setup

- Asthama dataset
  - 543 individuals, genotyped at 34 SNPs
  - Diploid data was transformed into 0/1 (for homozygotes) or 2 (for heterozygotes)
  - X=543x34 matrix
  - Y=Phenotype variable (continuous)
- A single phenotype was used for regression

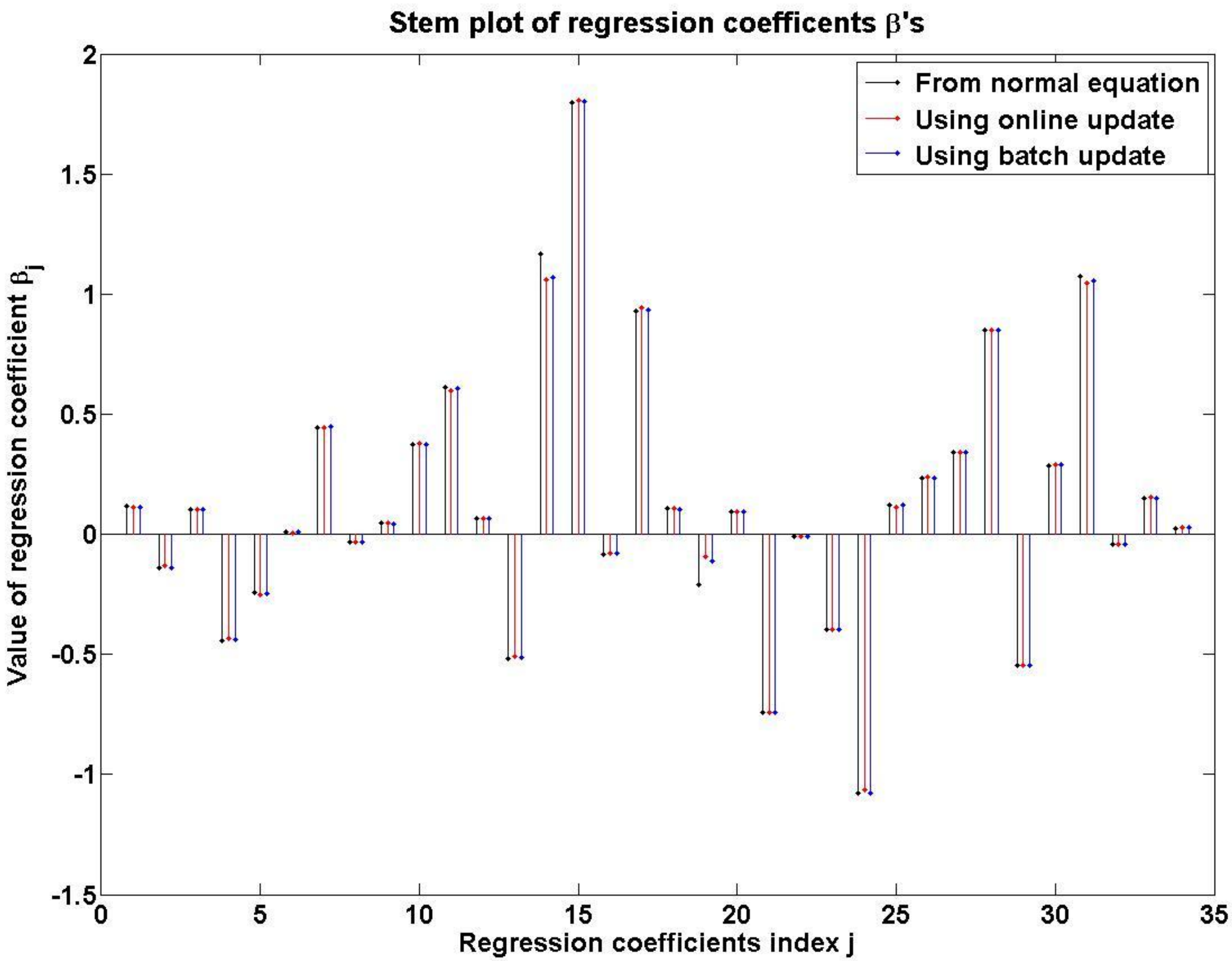
- Implementation details
  - Iterative methods: Batch update and online update implemented.
  - For both methods, step size α is chosen to be a small fixed value (10^{-6}). This choice is based on the data used for experiments.
  - Both methods are only run to a maximum of 2000 epochs or until the change in training MSE is less than 10-4

Convergence Curves

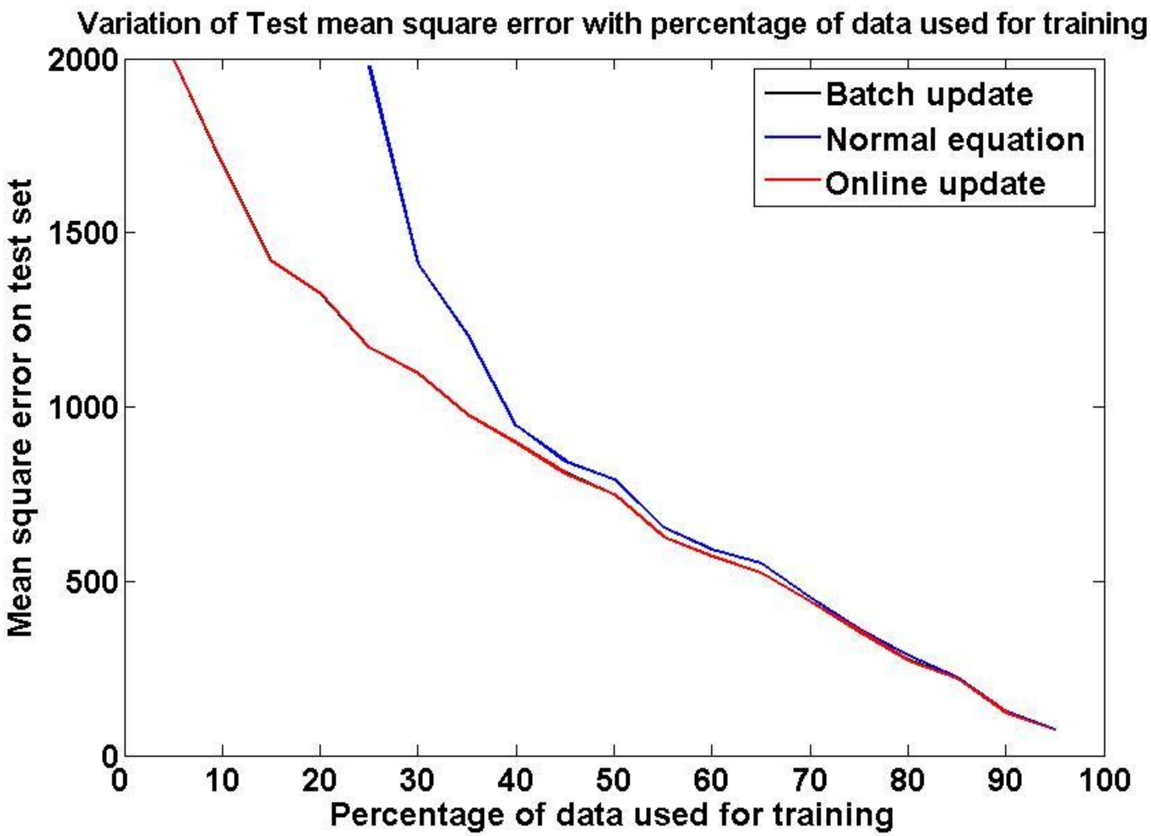


- For the batch method, the training MSE is initially large due to uninformed initialization
- In the online update, N updates for every epoch reduces MSE to a much smaller value.

The Learned Coefficients



Performance vs. Training Size



- The results from B and O update are almost identical. So the plots coincide.
- The test MSE from the normal equation is more than that of B and O during small training. This is probably due to overfitting.
- In B and O, since only 2000 iterations are allowed at most. This roughly acts as a mechanism that avoids overfitting.