Linear Regression

Pradeep Ravikumar

Co-instructor: Manuela Veloso

Machine Learning 10-701
Discrete to Continuous Labels

**Classification**

- **X = Document**
- **Y = Topic**
  - Sports
  - Science
  - News

- **X = Cell Image**
- **Y = Diagnosis**
  - Anemic cell
  - Healthy cell

**Regression**

- **X = Brain Scan**
- **Y = Age of a subject**
Regression Tasks

Estimating Energy Usage

Source EUI (kBtu/ft²)

Supermarket/ Grocery  | Hospital  | Bank Branch  | Medical Office  | Senior Care Facility  | Office  | Hotel  | Retail Store  | Courthouse  | Residence Hall/ Dormitory  | K-12 School  | Worship Facility  | Unrefrigerated Warehouse

X = building characteristics
Y = energy consumption

energystar.gov

Estimating Contamination

X = new location
Y = sensor reading
Performance Measures

**Performance Measure:** Quantifies knowledge gained

\[
\text{loss}(Y, f(X)) - \text{Measure of closeness between true label } Y \text{ and prediction } f(X)
\]

<table>
<thead>
<tr>
<th>$X$</th>
<th>Share price, $Y$</th>
<th>$f(X)$</th>
<th>loss($Y, f(X)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Past performance, trade volume etc. as of Sept 8, 2010</td>
<td>“$24.50”</td>
<td>“$24.50”</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>“$26.00”</td>
<td></td>
<td>1?</td>
</tr>
<tr>
<td></td>
<td>“$26.10”</td>
<td></td>
<td>2?</td>
</tr>
</tbody>
</table>

\[
\text{loss}(Y, f(X)) = (f(X) - Y)^2 \quad \text{square loss}
\]
Regression

Bayes optimal predictor:

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

\[ = \mathbb{E}[Y|X] \quad \text{(Conditional Mean)} \]
Regression

Optimal predictor:

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

Proof Strategy:

\[ R(f) \geq R(f^*) \text{ for any prediction rule } f \]

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

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

\[ = \mathbb{E} \left[ \frac{E[(f(X) - \mathbb{E}[Y|X])^2|X]}{E[(f(X) - \mathbb{E}[Y|X])|X]} \times 0 \right. \]

\[ + \frac{E[(E[Y|X] - Y)^2|X]}{E[(E[Y|X] - Y)^2|X]} \]

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

\[ \geq 0 \]
Regression

Optimal predictor:

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

\[ = \mathbb{E}[Y|X] \quad \text{(Conditional Mean)} \]

Model: Signal plus (zero-mean) Noise

\[ Y = f^*(X) + \epsilon \]

Model-based approach: estimate distribution \( P_{XY} \) and compute its conditional mean
Regression

Model-free approach: approximate response \( Y \) by function in function class (e.g. linear functions), without necessarily learning distribution of \( X \) and \( Y \)
Model-free approach: Empirical Risk Minimization

Optimal predictor:

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

Empirical Minimizer:

\[ \hat{f}_n = \arg \min_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \right) \]

Law of Large Numbers:

\[ \frac{1}{n} \sum_{i=1}^{n} [\text{loss}(Y_i, f(X_i))] \xrightarrow{n \to \infty} \mathbb{E}_{X,Y} [\text{loss}(Y, f(X))] \]
Restrict class of predictors

Optimal predictor:

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

Empirical Minimizer:

\[ \hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \]

Why?
Overfitting!
Empirical loss minimized by any function of the form

\[ f(x) = \begin{cases} Y_i, & x = X_i \text{ for } i = 1, \ldots, n \\ \text{any value,} & \text{otherwise} \end{cases} \]
Restrict class of predictors

Optimal predictor:

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

Empirical Minimizer:

\[ \hat{f}_n = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \]

Class of predictors

\[ \mathcal{F} \]
- Class of Linear functions
- Class of Polynomial functions
- Class of nonlinear functions
Regression algorithms

Training data \( \{(X_i, Y_i)\}_{i=1}^{n} \) \[\xrightarrow{\text{Learning algorithm}}\] Prediction rule \( \hat{f}_n \)

Linear Regression
Regularized Linear Regression – Ridge regression, Lasso
Polynomial Regression
Kernelized Ridge Regression
Gaussian Process Regression
Kernel regression, Regression Trees, Splines, Wavelet estimators, ...
Linear Regression

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

$\mathcal{F}_L$ - Class of Linear functions

Uni-variate case:

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

Multi-variate case:

$$f(X) = f(X^{(1)}, \ldots, X^{(p)}) = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \cdots + \beta_p X^{(p)}$$

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

Least Squares Estimator
Least Squares Estimator

\[ \hat{f}_n^L = \arg \min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \quad f(X_i) = X_i \beta \]

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

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

\[ A = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \cdots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \cdots & X_n^{(p)} \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \]
Least Squares Estimator

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

\[ J(\beta) = (A\beta - Y)^T (A\beta - Y) \]

\[ \left. \frac{\partial J(\beta)}{\partial \beta} \right|_{\hat{\beta}} = 0 \]
Least Square solution satisfies Normal Equations

\[(A^T A) \hat{\beta} = A^T Y\]

If \((A^T A)\) is invertible,

\[
\hat{\beta} = (A^T A)^{-1} A^T Y \quad \hat{f}^L_n(X) = X \hat{\beta}
\]
Example – linear regression

Linear Regression Relation Between Accidents & Population

Fatal traffic accidents per state vs. Population of state

Population of state × 10^7

0 0.5 1 1.5 2 2.5 3 3.5

0 500 1000 1500 2000 2500 3000 3500 4000 4500
Least Square solution satisfies Normal Equations

\[(A^T A) \hat{\beta} = A^T Y\]

If \( (A^T A) \) is invertible,

\[
\hat{\beta} = (A^T A)^{-1} A^T Y \quad \Rightarrow \quad f_n^L(X) = X \hat{\beta}
\]

When is \( (A^T A) \) invertible?
Recall: Full rank matrices are invertible. What is rank of \( (A^T A) \)?

\[
\text{Rank}(A^T A) = \text{number of non-zero eigenvalues of } (A^T A) = \text{number of non-zero singular values of } A \leq \min(n,p) \text{ since } A \text{ is } n \times p
\]

So, \( \text{rank}(A^T A) =: r \leq \min(n,p) \)
Not invertible if \( r < p \) (e.g. \( n < p \) i.e. high-dimensional setting)
Least Square solution satisfies Normal Equations

\[(A^T A)\beta = A^T Y\]

If \((A^T A)\) is invertible,

\[\hat{\beta} = (A^T A)^{-1} A^T Y\]

When is \((A^T A)\) invertible?
Recall: Full rank matrices are invertible. What is rank of \((A^T A)\)?

If \(A = USV^T\), then normal equations \((SV^T)\hat{\beta} = (U^TY)\)

\[r\text{ equations in } p\text{ unknowns. Under-determined if } r < p, \text{ hence no unique solution.}\]
Least Square solution satisfies Normal Equations

\[(A^T A)\hat{\beta} = A^T Y\]

If \((A^T A)\) is invertible,

\[
\hat{\beta} = (A^T A)^{-1} A^T Y
\]

\[
\hat{f}_n^L(X) = X\hat{\beta}
\]

When is \((A^T A)\) invertible?
Recall: Full rank matrices are invertible. What is rank of \((A^T A)\)?

What if \((A^T A)\) is not invertible?
Constrain solution i.e. Regularization (later)

Now: What if \((A^T A)\) is invertible but expensive (p very large)?
Optimization

Even when $A^T A$ is invertible, might be computationally expensive if $A$ is huge.

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

Treat as optimization problem

Observation: $J(\beta)$ is convex in $\beta$.

How to find the minimizer?
A function $l(w)$ is called **convex** if the line joining two points $l(w_1), l(w_2)$ on the function does not go below the function on the interval $[w_1, w_2]$

(Strictly) Convex functions have a unique minimum!
Optimizing convex functions

- Minimum of a convex function can be reached by the Gradient Descent Algorithm.

**Gradient Descent Algorithm**

Initialize: Pick \( w \) at random

Gradient:

\[
\nabla_w l(w) = \left[ \frac{\partial l(w)}{\partial w_0}, \ldots, \frac{\partial l(w)}{\partial w_p} \right]'
\]

Update rule:

\[
\Delta w = \eta \nabla_w l(w)
\]

\[
w_{i}(t+1) \leftarrow w_i(t) - \eta \left. \frac{\partial l(w)}{\partial w_i} \right|_t
\]

Learning rate, \( \eta > 0 \)
Gradient Descent

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

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

Since \(J(\beta)\) is convex, move along negative of gradient

Initialize: \(\beta^0\)

Update:

\[
\beta^{t+1} = \beta^t - \alpha \frac{\partial J(\beta)}{\partial \beta} \bigg|_t \\
= \beta^t - \alpha A^T (A\beta^t - Y)
\]

0 if \(\hat{\beta} = \beta^t\)

Stop: when some criterion met e.g. fixed # iterations, or \(\left.\frac{\partial J(\beta)}{\partial \beta}\right|_{\beta^t} < \varepsilon\).
Effect of step-size $\alpha$

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

Small $\alpha$ => Slow convergence but small residual error
Regularized Least Squares

What if \((A^T A)\) is not invertible?

r equations, p unknowns – underdetermined system of linear equations
many feasible solutions
Need to constrain solution further

e.g. bias solution to “small” values of \(\beta\) (small changes in input don’t
translate to large changes in output)

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

\[
= \arg \min_{\beta} \quad (A\beta - Y)^T (A\beta - Y) + \lambda \|\beta\|^2_2
\]

\[
\hat{\beta}_{\text{MAP}} = (A^T A + \lambda I)^{-1} A^T Y
\]

Is \((A^T A + \lambda I)\) invertible?
Understanding regularized Least Squares

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

Ridge Regression:
\[
\text{pen}(\beta) = \|\beta\|^2_2
\]

\(\beta_s\) with constant \(J(\beta)\)
(level sets of \(J(\beta)\))

\(\beta_s\) with constant \(l2\) norm
(level sets of \(\text{pen}(\beta)\))

Unregularized Least Squares solution
Regularized Least Squares

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

$r$ equations, $p$ unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to “small” values of $b$ (small changes in input don’t translate to large changes in output)

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

Ridge Regression
(l2 penalty)

$$\lambda \geq 0$$

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

Lasso
(l1 penalty)

Many $b$ can be zero – many inputs are irrelevant to prediction in high-dimensional settings (typically intercept term not penalized)
Regularized Least Squares

What if \((A^T A)\) is not invertible?

\(r\) equations, \(p\) unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to “small” values of \(\beta\) (small changes in input don’t translate to large changes in output)

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

Ridge Regression
(l2 penalty)

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

Lasso
(l1 penalty)

No closed form solution, but can optimize using sub-gradient descent (packages available)
Ridge Regression vs Lasso

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

Ridge Regression:
\[
\text{pen}(\beta) = \|\beta\|_2^2
\]

Lasso:
\[
\text{pen}(\beta) = \|\beta\|_1
\]

Ideally \(l_0\) penalty, but optimization becomes non-convex

\(\beta\)s with constant \(J(\beta)\) (level sets of \(J(\beta)\))

\(\beta\)s with constant \(l_2\) norm

\(\beta\)s with constant \(l_1\) norm

\(\beta\)s with constant \(l_0\) norm

Lasso (\(l_1\) penalty) results in sparse solutions – vector with more zero coordinates
Good for high-dimensional problems – don’t have to store all coordinates, interpretable solution!
Lasso vs Ridge

Lasso Coefficients

Ridge Coefficients
Regularized Least Squares – connection to MLE and MAP (Model-based approaches)
Least Squares and M(C)LE

Intuition: Signal plus (zero-mean) Noise model

\[ Y = f^*(X) + \epsilon = X\beta^* + \epsilon \]

\[ \epsilon \sim \mathcal{N}(0, \sigma^2 I) \quad Y \sim \mathcal{N}(X\beta^*, \sigma^2 I) \]

\[ \hat{\beta}_{\text{MLE}} = \arg \max_{\beta} \log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n) \]

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

Least Square Estimate is same as Maximum Conditional Likelihood Estimate under a Gaussian model!
Regularized Least Squares and M(C)AP

What if \((A^T A)\) is not invertible?

\[
\hat{\beta}_{\text{MAP}} = \arg\max_{\beta} \log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n) + \log p(\beta)
\]

Conditional log likelihood log prior

I) Gaussian Prior

\[
\beta \sim \mathcal{N}(0, \tau^2 I) \quad \quad p(\beta) \propto e^{-\beta^T \beta / 2\tau^2}
\]

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

\[
\hat{\beta}_{\text{MAP}} = (A^T A + \lambda I)^{-1} A^T Y
\]

Ridge Regression
Regularized Least Squares and M(C)AP

What if \((A^T A)\) is not invertible?

\[
\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n) + \log p(\beta)
\]

<table>
<thead>
<tr>
<th>Conditional log likelihood</th>
<th>log prior</th>
</tr>
</thead>
</table>

I) Gaussian Prior

\[
\beta \sim \mathcal{N}(0, \tau^2 I) \quad p(\beta) \propto e^{-\beta^T \beta / 2\tau^2}
\]

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

Ridge Regression

Prior belief that \(\beta\) is Gaussian with zero-mean biases solution to “small” \(\beta\)
Regularized Least Squares and M(C)AP

What if \((A^T A)\) is not invertible?

\[
\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n) + \log p(\beta)
\]

\text{Conditional log likelihood} \quad \text{log prior}

II) Laplace Prior

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

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

\text{Lasso}

Prior belief that \(\beta\) is Laplace with zero-mean biases solution to "sparse" \(\beta\)