Variations on Regression Models

Prof. Bennett
Math Models of Data Science
2/02/06
Outline

- Steps in modeling
- Review of Least Squares model
- Model in E & K pg 24-29
- Aqualsol version of E&K
- Other loss functions
- Other regularization
Modeling Process

- Gather data: input and output pairs
- Represent data mathematically: \((X,Y)\)
- Select parametric model: \(g(x)\)
- Select loss function: \(\text{Loss}\)
- Select regularization
- Optimize parameters with respect to given loss
- Estimate out-of-sample accuracy
Predict Drug Bioavailability

Aqua solubility = Aquasol
525 descriptors generated
- Electronic TAE
- Traditional
197 molecules with tested solubility

\[ y \in \mathbb{R} \]

\[ x_i \in \mathbb{R}^{525} \]

\[ \ell = 197 \]
Modeling Process

- Gather data: input and output pairs
- Represent data mathematically: \((X,Y)\)
- Select parametric model: \(g(x)\)
- Select loss function: Loss
- Select regularization
- Optimize parameters with respect to given loss
- Estimate out-of-sample accuracy
1-d Regression with bias

\[ \langle w, x \rangle + b \]
Modeling Process

- Gather data: input and output pairs
- Represent data mathematically: \((X,Y)\)
- Select parametric model: \(g(x)\)
- Select loss function: Loss
- Select regularization
- Optimize parameters with respect to given loss
- Estimate out-of-sample accuracy
Linear Regression

Given training data:

\[ S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_\ell, y_\ell) \} \]

points \( x_i \in \mathbb{R}^n \) and labels \( y_i \in \mathbb{R} \)

_construct linear function:\n
\[ g(x) = \langle x, w \rangle = x'w = \sum_{i=1}^{n} w_i (x)_i + b \]

Goal for future data \((x, y)\) with \( y \) unknown

\[ g(x) \approx y \]
Least Squares Approximation

- Want $g(x) \approx y$

- Define error

  $f(x, y) = y - g(x) = \xi$

- Minimize loss

  $$L(g, s) = \sum_{i=1}^{\ell} (y_i - g(x_i))^2$$
Least Squares Loss

\[ L(w) = \sum_{i=1}^{\ell} (x_i 'w + b - y_i)^2 \]

\[ = \|Xw + eb - y\|^2 \quad 2-norm \]

\[ = (Xw + eb - y)'(Xw + eb - y) \]
Convex Functions

A function $f$ is (strictly) convex on a convex set $S$, if and only if for any $x, y \in S$,

$$f(\lambda x + (1-\lambda)y)(<) \leq \lambda f(x) + (1-\lambda)f(y)$$

for all $0 \leq \lambda \leq 1$. 
Theorem

- Consider problem \( \min f(x) \) unconstrained.
- If \( \nabla f(\bar{x}) = 0 \) and \( f \) is convex, then \( \bar{x} \) is a global minimum.

**Proof:**

\[
\forall y \quad f(y) \geq f(\bar{x}) + (y - \bar{x})' \nabla f(\bar{x}) \quad \text{by convexity of } f
\]

\[
= f(\bar{x}) \quad \text{since } \nabla f(\bar{x}) = 0.
\]
Stationary Points

Note that this condition is not sufficient

$$\nabla f (x^*) = 0$$

Also true for local max and saddle points
Modeling Process

- Gather data: input and output pairs
- Represent data mathematically: \((X,Y)\)
- Select parametric model: \(g(x)\)
- Select loss function: \(\text{Loss}\)
- Select regularization
- Optimize parameters with respect to given loss
- Estimate out-of-sample accuracy
Optimal Solution

- **Want:** \( y \approx Xw + be \)  \( e \) is a vector of ones

- **Mathematical Model:**

\[
\min_w \ L(w, b, S) = \| y - (Xw + eb) \|^2 + \| w \|^2
\]

- **Optimality Conditions:**

\[
\frac{\partial L(w, b, S)}{\partial w} = 2X'(y - Xw - eb) + 2w = 0
\]

\[
\frac{\partial L(w, b, S)}{\partial b} = 2e'(y - Xw - eb) = 0
\]
Thus:

\[ e'eb = e'y - e'Xw \]

\[ \Rightarrow b = \frac{e'y}{\ell} - \frac{e'Xw}{\ell} = mean(y) - mean(X)'w \]

\[ (X'X + \lambda I)w = X'y - X'eb \]

Idea: Scale data so means are 0, e.g.

\[ e'y = 0 \]
\[ e'X = 0 \]
Recenter Data

- Shift $y$ by mean

\[ \mu = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i \quad y_i := y_i - \mu \]

- Shift $x$ by mean

\[ \bar{x} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i \quad x_i := x_i - \bar{x} \]
Ridge Regression with bias

Center data by $\bar{x}$ and $\mu$

$X = X - e\mu'$

$y = y - \mu e$

Calculate $w$

$$w = (X'X + \lambda I)^{-1}X'y$$

Calculate $b$

$$b = (\mu - \bar{x}'w)$$

To predict new point $g(x) = x'w + b$
Modeling Process

- Gather data: input and output pairs
- Represent data mathematically: (X,Y)
- Select parametric model: \( g(x) \)
- Select loss function: \( \text{Loss} \)
- Select regularization
- Optimize parameters with respect to given loss
- Estimate out-of-sample accuracy
Generalization

To estimate generalization error:
Divide test into training set = Xtrain
100 points in Aquasol
and test set = Xtest
97 points in Aquasol
Create g(x) using Xtrain
Evaluate on Xtest
Train and Test for $\lambda$
Could we do better

- Other loss functions?
- Other types of regularization?
Nonlinear time series

Read E&K pg 24-25

What are the modeling tricks?
Linear Programming

Matlab can solve linear programs of the following form for given \( f, A, A_{eq}, lb, ub \)

\[
\begin{align*}
\min_x & \quad f'x \\
A \cdot x & \leq b \\
A_{eq} \cdot x & = b_{eq} \\
lb & \leq x \leq ub
\end{align*}
\]
Sample Problem

\[ T = [1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \ 14]' \]

\[ R = [-1.7500 \ -4.0000 \ -5.7500 \ -7.0000 \ -7.7500 \ -8.0000 \ -7.7500 \ -7.0000 \ -5.7500 \ -4.0000 \ -1.7500 \ 1.0000 \ 4.2500 \ 8.0000] \]
Let $i$th row of $B$ be $\begin{bmatrix} t_i^2 & t_i & 1 \end{bmatrix}$

Problem becomes

$$\min_{w,a,b,c} \begin{bmatrix} 1000 \end{bmatrix}^{'} \begin{bmatrix} w & a & b & c \end{bmatrix}^{'}$$

$$\begin{bmatrix} -e & -B \\ -e & B \end{bmatrix} \begin{bmatrix} w \\ a \\ b \\ c \end{bmatrix} \leq \begin{bmatrix} -R \\ +R \end{bmatrix}$$
Comparison of norms \( ||w||=1 \)
Inconsistent System of equations

Want $Ax \approx b$?
How does this correspond to our framework?
What are the loss functions?
Review of norms

2-norm

$$\| y \|_2 = \sqrt{\sum_{i=1}^{n} y_i^2}$$

1-norm

$$\| y \|_1 = \sum_{i=1}^{n} | y_i |$$

∞-norm

$$\| y \|_\infty = \max_{i} | y_i |$$
Translate the ideas to our framework

Using the ideas in E&K, formulate the following problem as a linear program

$$\min_{w,b} \|Xw + b - y\|_\infty$$
Translate the ideas to our framework

Using the ideas in E&K, formulate the following problem as a linear program

$$\min_{w,b} \left\| Xw + b - y \right\|_1$$
Translate the ideas to our framework

Using the ideas in E&K, formulate the following problem as a linear program

\[ \min_{w,b} \left\| Xw + b - y \right\|_1 + \lambda \left\| w \right\|_1 \]
Challenge problem

Popular loss function

ε-insensitive

\[
\min_{w,b} \sum_i \max(\{ |x_i'w + b - y| - \varepsilon, 0 \})
\]
Discussion Questions

- Sketch each type of loss function and discuss advantages and disadvantages of different loss functions? (1-norm, 2-norm, $\infty$-norm, $\varepsilon$-insensitive)

- What would be the ramification of using these same functions for regularization?