Regularized Linear Regression

Nakul Gopalan
Georgia Tech

These slides are adopted based on slides from Andrew Zisserman, Jonathan Taylor, Chao Zhang, Mahdi Roozbahani and Yaser Abu-Mostafa.
Recap

- Linear regression:
- \( Y = \theta X \)
- MSE
Polynomial regression
Polynomial regression when order not known
Outline

• Overfitting and regularized learning
• Ridge regression
• Lasso regression
• Determining regularization strength
Regression: Recap

- Suppose we are given a training set of N observations 

\[(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N)\]

- Regression problem is to estimate \(y(x)\) from this data
Regression: Recap

- Want to fit a polynomial regression model

\[ y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d + \epsilon \]

- \( z = \{1, x, x^2, ..., x^d\} \in \mathbb{R}^d \) and \( \theta = (\theta_0, \theta_1, \theta_2, ..., \theta_d)^T \)

\[ y = z\theta \]
Real Regression Problem !!!
Which One is Better?

Can we increase the maximal polynomial degree to very large, such that the curve passes through all training points?

No, this can lead to overfitting!
The Overfitting Problem

- The training error is very low, but the error on test set is large.
- The model captures not only patterns but also noisy nuisances in the training data.
The Overfitting Problem

• In regression, overfitting is often associated with large Weights (severe oscillation)

• How can we address overfitting?
Regularization
(smart way to cure overfitting disease)

Fit a linear line on sinusoidal with just two data points
Who is the winner?

\( \bar{g}(x) \): average over all lines

**without regularization**

- bias = 0.21; var = 1.69

**with regularization**

- bias = 0.23; var = 0.33
Regularized Learning

Minimize

$$E(\theta) + \frac{\lambda}{N} \theta^T \theta$$

• Cost function – squared loss:

$$\tilde{E}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\{ f(x_i, \theta) - y_i \right\}^2 + \frac{\lambda}{N} \|\theta\|^2$$

Why this term leads to regularization of parameters
Polynomial Model

Want to fit a polynomial regression model

\[ y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d + \epsilon \]

Let’s rewrite it as:

\[ y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_d + \epsilon = z\theta \]
Regularizing is just constraining the weights ($\theta$)

For example: let’s do a **hard** constraining

$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_d$$

subject to

$$\theta_d = 0 \text{ for } d > 2$$

$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + 0 + \cdots + 0$$
Let’s not penalize $\theta$ in such a harsh way
let’s cut them some slack

$$\theta = \arg\min_\theta E(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y^i - z_i \theta)^2$$

Minimize $\frac{1}{N} (z\theta - y)^T (z\theta - y)$

Subject to $\theta^T\theta \leq C$

For simplicity let’s call $\theta_{lin}$ as weights’ solution for non constrained one
and $\theta$ for the constrained model.
$$E(\theta) = \frac{1}{N} (z\theta - y)^T (z\theta - y)$$

Possible graph for $E(\theta)$ for different values of $\theta_0$ and $\theta_1$ and given observation data

3D view

Top view
Gradient of $\theta^t \theta$

\[
\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad \Rightarrow \quad \theta^t \theta = \theta_0^2 + \theta_1^2
\]

If you imagine standing at a point $(\theta_0, \theta_1)$, $\nabla(\theta^T \theta)$ tells you which direction you should travel to increase the value of $\theta^T \theta$ most rapidly.

\[
\nabla(\theta^T \theta) = \begin{bmatrix}
\frac{\partial}{\partial (\theta_0)} (\theta^T \theta) \\
\frac{\partial}{\partial (\theta_1)} (\theta^T \theta)
\end{bmatrix} = \begin{bmatrix} 2\theta_0 \\ 2\theta_1 \end{bmatrix} \approx \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}
\]

$\nabla(\theta^T \theta)$ is a vector field any line passing through the center of the circle.
Plotting the regularization term $\theta^t \theta$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$  \Rightarrow  $\theta^t \theta = \theta_0^2 + \theta_1^2$

3D view

Top view
\[ E(\theta) = \frac{1}{N} (z\theta - y)^T (z\theta - y) \]

Subject to \( \theta^t \theta \leq C \)

Find a solution in \( \theta^t \theta \) that minimizes \( E(\theta) \)

\( \theta_{lin} \) is the solution (min absolute)

\( E(\theta) \): which is constant on the surface of the ellipsoid
Constraint and Loss

\[ \theta_0 \]

\[ \theta_1 \]

\[ \theta_{\text{lin}} \]
Considering the below $E(\theta)$ and $C$ what is a $\theta$ candidate here?

$\nabla E$: the gradient (rate) in objective function which minimizes error (orthogonal to ellipse. Changes happen in orthogonal direction)

$\theta^t \theta = Constraint = C$

What is the orthogonal direction on the other surface?

It is just $\theta$, a line passing through center of the circle

Applying a constrain $\theta^t \theta$, where the best solution happens?

On the boundary of the circle, as it is the closest one to the minimum absolute
Considering the below $E(\theta)$ and $C$ what is the best $\theta$ solution here?

$\nabla E(\theta) \propto -\theta$

$\nabla E(\theta) = -\frac{2\lambda}{N}\theta$

$\nabla E(\theta) + 2\frac{\lambda}{N}\theta = 0$

Let's do integration

Minimize $E(\theta) + \frac{\lambda}{N}\theta^T\theta$

$C \uparrow \lambda \downarrow$
Outline

• Overfitting and regularized learning
• Ridge regression
• Lasso regression
• Determining regularization strength
Ridge Regression

- Cost function – squared loss:

\[ \tilde{E}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\{ f(x_i, \theta) - y_i \right\}^2 + \frac{\lambda}{N} ||\theta||^2 \]

- Regression function for x (1D):

\[ f(x, \theta) = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_d + \epsilon = z\theta \]
Solving for the Weights $\theta$

Notation: write the target and regressed values as $N$-vectors

\[
y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_N \end{pmatrix} \quad f = \begin{pmatrix} z(x_1)\theta \\ z(x_2)\theta \\ \cdot \\ \cdot \\ z(x_N)\theta \end{pmatrix} = z\theta = \begin{pmatrix} 1 & z_1(x_1) & \cdots & z_d(x_1) \\ 1 & z_1(x_2) & \cdots & z_d(x_2) \\ \cdot \\ \cdot \\ 1 & z_1(x_N) & \cdots & z_d(x_N) \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \cdot \\ \cdot \\ \theta_d \end{pmatrix}
\]

$z$ is an $N \times D$ design matrix

e.g. for polynomial regression with basis functions up to $x^2$

\[
z\theta = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdot & \cdot & \cdot \\ 1 & x_N & x_N^2 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}
\]
\[ \tilde{E}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \{f(x_i, \theta) - y_i\}^2 + \frac{\lambda}{N} \|\theta\|^2 \]

\[ = \frac{1}{N} \sum_{i=1}^{N} (y_i - z_i \theta)^2 + \frac{\lambda}{N} \|\theta\|^2 \]

\[ = \frac{1}{N} (y_i - z\theta)^2 + \frac{\lambda}{N} \|\theta\|^2 \]

Now, compute where derivative w.r.t. \( \theta \) is zero for minimum

\[ \frac{\tilde{E}(\theta)}{d\theta} = -z^T (y - z\theta) + \lambda \theta \]

Hence

\[ (z^T z + \lambda I)\theta = z^T y \]

\[ \theta = (z^T z + \lambda I)^{-1} z^T y \]
D basis functions, N data points

\[
\theta = (z^T z + \lambda I)^{-1} z^T y
\]

\[
\begin{bmatrix}
\theta
\end{bmatrix} = \begin{bmatrix}
\theta
\end{bmatrix}
\]

assume \(N > D\)

\[
\begin{bmatrix}
Dx1 & DxD & DxN & Nx1
\end{bmatrix}
\]

- This shows that there is a unique solution.

- If \(\lambda = 0\) (no regularization), then

\[
\theta = (z^T z)^{-1} z^T y = z^+ y
\]

where \(z^+\) is the pseudo-inverse of \(z\) (\texttt{pinv} in Matlab)

- Adding the term \(\lambda I\) improves the \textbf{conditioning} of the inverse, since if \(Z\) is not full rank, then \((z^T z + \lambda I)\) will be (for sufficiently large \(\lambda\))

- As \(\lambda \to \infty\), \(\theta \to \frac{1}{\lambda} z^T y \to 0\)
Ridge Regression Example

• The red curve is the true function (which is not a polynomial)

• The data points are samples from the curve with added noise in y.

• There is a choice in both the degree, D, of the basis functions used, and in the strength of the regularization

\[
f(x, \theta) = z\theta \quad z: x \rightarrow z \quad \mathbb{R} \rightarrow \mathbb{R}^{D+1}
\]

\[
\tilde{E}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\{f(x_i, \theta) - y_i\right\}^2 + \frac{\lambda}{N} \|\theta\|^2
\]

\(\theta\) is a \(D+1\) dimensional vector
N = 9 samples, D = 7
Outline

• Overfitting and regularized learning
• Ridge regression
• Lasso regression
• Determining regularization strength
Regularized Regression

Minimize with respect to $\theta$

$$\sum_{i=1}^{N} l(f(x_i, \theta), y_i) + \lambda R(\theta)$$

- loss function
- regularization

- There is a choice of both loss functions and regularization
- So far we have seen – “ridge” regression
  - squared loss: $$\sum_{i=1}^{N} (y_i - f(x_i, \theta))^2$$
  - squared regularizer: $$\lambda \| \theta \|^2$$

Now let's look at another regularization choice.
The Lasso Regularization (norm one)

- LASSO = Least Absolute Shrinkage and Selection

Minimize with respect to $\theta$

$$\sum_{i=1}^{N} l(f(x_i, \theta), y_i) + \lambda R(\theta)$$

- loss function
- regularization

- This is a quadratic optimization problem
- There is a unique solution
- $p$-Norm definition: $\|\theta\|_p = \left( \sum_{j=1}^{d} |\theta_i|^p \right)^{1/p}$
Let’s say we have two parameters ($\theta_0$ and $\theta_1$)

Minimize $E(\theta) = \frac{1}{N} (zw - y)^T (z\theta - y)$

Subject to $\theta \leq C$

$\theta = \begin{bmatrix} \theta_0 \\ 0 \end{bmatrix}$

Interesting way for feature selection

$E(\theta)$: which is constant on the surface of the ellipsoid

Sharp edges

C
Outline

• Overfitting and regularized learning
• Ridge regression
• Lasso regression
• Determining regularization strength
Leave-One-Out Cross Validation

For every \( i = 1, \ldots, n \):

- train the model on every point except \( i \),
- compute the test error on the held out point.

Average the test errors.

\[
\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2
\]
K-Fold Cross Validation

Split the data into $k$ subsets or *folds*.

For every $i = 1, \ldots, k$:

- train the model on every fold except the $i$th fold,
- compute the test error on the $i$th fold.

Average the test errors.
Choosing $\lambda$ Using Validation Dataset

Pick up the lambda with the lowest mean value of rmse calculated by Cross Validation approach
Take-Home Messages

• What is overfitting
• What is regularization
• How does Ridge regression work
• Sparsity properties of Lasso regression
• How to choose the regularization coefficient $\lambda$