Regularization

Sanjay Lall and Stephen Boyd

EE104
Stanford University
Sensitivity

- we have a linear predictor \( \hat{y} = g(x) = \theta^T x \)

- if \( |\theta_i| \) is large, then the prediction is very sensitive to \( x_i \) 
  (i.e., small changes in \( x_i \) lead to large changes in the prediction)

- large sensitivity can lead to overfit, poor generalization 
  (which would turn up in validation)

- for \( x_1 = 1 \) (the constant feature), there is no sensitivity, since the feature 
  does not change

- suggests that we would like \( \theta \) (or \( \theta_{2:d} \) if \( x_1 = 1 \)) not too large
we will measure the size of $\theta$ using a *regularizer* function $r : \mathbb{R}^d \to \mathbb{R}$

$r(\theta)$ is a measure of the size of $\theta$ (or $\theta_{2:d}$)

- **quadratic regularizer** (a.k.a. $\ell_2$ or sum-of-squares):
  
  $$r(\theta) = ||\theta||^2 = \theta_1^2 + \cdots + \theta_d^2$$

- **absolute value regularizer** (a.k.a. $\ell_1$):
  
  $$r(\theta) = ||\theta||_1 = |\theta_1| + \cdots + |\theta_d|$$
Regularized empirical risk minimization

- our model should fit the given data well, i.e., loss

\[ L(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta^\top x^i, y^i) \]

should be small

- the model should not be too sensitive, i.e., \( r(\theta) \) should be small

- \textit{regularized empirical risk minimization}: choose \( \theta \) to minimize weighted sum

\[ L(\theta) + \lambda r(\theta) \]

where \( \lambda \geq 0 \) is the \textit{regularization parameter} (or \textit{hyper-parameter})

- an optimization problem
Regularized empirical risk minimization

- for $\lambda = 0$, reduces to ERM
- RERM produces a *family* of models, one for each value of $\lambda$
- in practice, one choose a few tens of values of $\lambda$, usually logarithmically spaced over a wide range
- use validation to choose among the candidate models
- we choose the largest value of $\lambda$ that gives near minimum test loss (*i.e.*, least sensitive model that generalizes well)
Ridge regression

- **ridge regression**: square loss and regularizer \( r(\theta) = ||\theta||^2 \) (or \( ||\theta_{2:d}||^2 \) if \( x_1 = 1 \))
- also called *Tykhonov regularized least squares*

- weighted sum objective function is

\[
L(\theta) + \lambda r(\theta) = ||X\theta - y||^2 + \lambda||\theta||^2
\]

\[
= \left\| \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix} \theta - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2
\]

- so optimal \( \theta \) is

\[
\theta^* = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix}^\dagger \begin{bmatrix} y \\ 0 \end{bmatrix} = (X^TX + \lambda I)^{-1} X^T y
\]

- (how do you modify this to handle \( r(\theta) = ||\theta_{2:d}||^2 \)?)
Example: House prices

- sales prices of 2930 homes in Ames, Iowa from 2006 to 2010
- 80 features
- we use 16 features
Example: Regression

we manually remove 4 outliers with area > 4000
(we’ll see later how to detect outliers)
Example: Regression

- split data randomly into 1164 training, 291 test
- target is $\log(\text{price})$
- standardize all features (and $\log(\text{price})$)
- training loss 0.1060, test loss 0.1361
- plot shows all test points
Example: Ridge regression

- leftmost loss is training loss with no regularization: 0.1060
- rightmost loss is variance of training data: 0.9787
- plot of $\theta_i$ versus $\lambda$ (on right) is called **regularization path**
- rightmost $\theta$ has $\theta_0 = -0.0043$, the mean of training $y$ values
Example: Ridge regression

- Regularization $\lambda = 187$ is optimal; improves test performance a bit
- $\theta$ is shrunk by regularization, so model is less sensitive
Example: Ridge regression

- least squares test loss is 0.1361, with $||\theta|| \approx 0.55$
- ridge regression test loss (with $\lambda = 178$) is 0.1295 with $||\theta|| \approx 0.46$
- ridge regression model is less sensitive
Example: Piecewise linear fit

features $x = (1, u, (u - 0.2)_+, (u - 0.4)_+, (u - 0.6)_+, (u - 0.8)_+)$

$\lambda = 1$ gives $\theta = (0.36, 0.25, -0.057, -0.056, 0.089, 0.26)$

$\lambda = 10^{-5}$ gives $\theta = (0.05, 2.9, -3.9, 1.6, -2, 4.8)$
Fitting models with more parameters than data points

- this makes no sense in general
- but with regularization, you can do this
  - $\lambda = 1$ gives $\theta = (0.55, 0.039, 0.033, 0.022, 0.011, -0.0007)$
  - $\lambda = 10^{-5}$ gives $\theta = (0.46, 0.42, 0.22, -0.18, -0.58, -0.98)$
Fitting models with more parameters than data points

- minimum point balances fitting training data versus sensitivity