INTRODUCTION TO MACHINE LEARNING

REGULARIZATION ON LINEAR MODEL

* Some contents are adapted from Dr. Hung Huang and Dr. Chengkai Li at UT Arlington

Mingon Kang, Ph.D.

Department of Computer Science @ UNLV

Motivation

- If more than two independent variables are highly correlated:

```
> x1 <- rnorm(20);x2 <- rnorm(20,mean=x1,sd=.01)
> cor(x1,x2)
[1] 0.9999423
> y <- rnorm(20,mean=3+x1+x2)
> coef(lm(y~x1+x2))
(Intercept)          x1          x2
   2.582064   39.971344  -38.040040
```

- The intercept is approximated well, but coefficients?

Reference: http://web.as.uky.edu/statistics/users/pbreheny/603/2-20.pdf

Motivation

- It happens because x1 and x2 are highly correlated.
  - RSS(40, -38) = 21.7 (our estimate) is very closed to RSS(1, 1) = 22.6 (the truth)
- Effective way of dealing with this problem is through penalization:
  - Instead of minimizing RSS only, we consider an additional term in the regression form…

Reference: http://web.as.uky.edu/statistics/users/pbreheny/603/2-20.pdf

Ridge Regression

- Ridge Regression Model

$$Minimize \sum_{i=1}^{n}(y_i - \mathbf{Xb})^2$$

$$s.t. \sum_{j=1}^{p} {b_j}^2 \leq c$$

Ridge Regression

- Why does this help?
  - Smaller coefficients give less sensitivity of the variables.

```
> coef(lm(y~x1+x2))
(Intercept)          x1          x2
   2.582064   39.971344  -38.040040
```

```
> lm.ridge(y~x1+x2,lambda=1)
                 x1        x2
2.6214998 0.9906773 0.8973912
```

Reference: http://web.as.uky.edu/statistics/users/pbreheny/603/2-20.pdf

Ridge Regression

- Lagrange Multiplier
  - A strategy for finding the local maxima or minima of a function subject to equality/inequality constraints

Minimizing

$$\sum_{i=1}^{n} f(x) \; s.t. \, g(x) \leq C$$

Equivalent to minimizing

$$\sum_{i=1}^{n} f(x) + \lambda g(x),$$

Where λ is positive.

Ridge Regression

- Ridge Regression Model

$$Minimize \sum_{i=1}^{n}(y_i - \mathbf{Xb})^2 + \lambda\|\mathbf{b}\|^2,$$

where $\|\mathbf{b}\|^2$ is L-2 norm of **b** (Euclidean distance)

P-norm ($p \geq 1$)

$$\|x\|_{\boldsymbol{p}} := \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{x_i}|^{\boldsymbol{p}}\right)^{\boldsymbol{1/p}}$$

p=1, Manhattan norm (L-1 norm); p=2, Euclidean norm; p=∞, maximum norm

Optimization

$$\begin{aligned}\mathrm{H}(\mathbf{b},\lambda) &= (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})+\lambda\mathbf{b}'\mathbf{b}\\ &= \mathbf{y}'\mathbf{y}-\mathbf{2b}'\mathbf{X}'\mathbf{y}+\mathbf{b}'\mathbf{X}'\mathbf{Xb}+\lambda\mathbf{b}'\mathbf{b}\end{aligned}$$

$$\frac{\partial \mathrm{H}(\mathbf{b},\lambda)}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{Xb} + 2\lambda\mathbf{b} = \mathbf{0}$$

$$(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

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

$\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}$ is always invertible. Always gives a unique solution, $\hat{\mathbf{b}}$

Ridge Regression

- Similar to the ordinary least squares solution, but with the addition of a "ridge" regularization
  - $\lambda \rightarrow 0$, $\hat{\mathbf{b}}^{ridge} \rightarrow \hat{\mathbf{b}}^{OLS}$
  - $\lambda \rightarrow \infty$, $\hat{\mathbf{b}}^{ridge} \rightarrow 0$

- Applying the ridge regression penalty has the effect of shrinking the estimates toward zero
- Introduce bias but reduce the variance of the estimate