Regularization

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Andrew Ng (Stanford), Eric Eaton (UPenn), David Kauchak (Pomona), and the many others who made their course materials freely available online.

Regularization

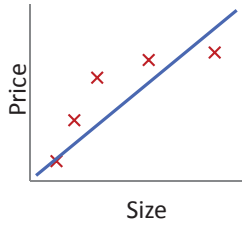
Learning Goals

- Describe goal of regularization
- Describe how to solve regularized regression using gradient descent and normal equations
- Describe common regularizers and tradeoffs

Generalization Error

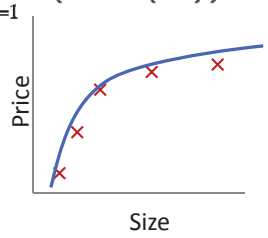
How are training error and generalization error related?

empirical risk $R_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L\left(\boldsymbol{y}^{(i)}, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)$ risk $R(\boldsymbol{\theta}) = \mathbb{E}\left[L\left(\boldsymbol{y}^{(i)}, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)\right]$



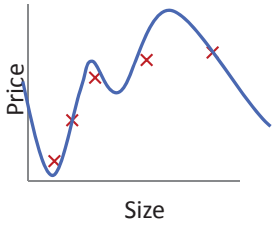
$\theta_0 + \theta_1 x$

underfitting
(high bias)



$\theta_0 + \theta_1 x + \theta_2 x^2$

correct fit



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

overfitting
(high variance)

structural error
hypothesis space cannot model true relationship
$\Rightarrow$ more data does not help
$\Rightarrow$ need larger $\mathcal{H}$

balance $\Leftrightarrow$

estimation (approximation) error
hypothesis space can model true relationship BUT hard to identify correct model due to large $|\mathcal{H}|$, small n, or noise
$\Rightarrow$ reduce $\mathcal{H}$
$\Rightarrow$ add more data

Based on slide by Eric Eaton [example by Andrew Ng]

Regularization

What if …

- we have a limited # of training examples ($n < d$), or
- we want to automatically control the complexity of the learned hypothesis?

Idea: penalize large values of θ_j

Why prefer small weights?

Regularized Linear Regression

Incorporate into cost function

$$J(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2}_{\text{model fit to data}} + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2}_{\text{regularization}}$$

- $\lambda \geq 0$ is regularization parameter – trade off between
  - small λ
  - large λ
- Note that θ_0 is not regularized!

Based on slide by Eric Eaton

Understanding Regularization

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

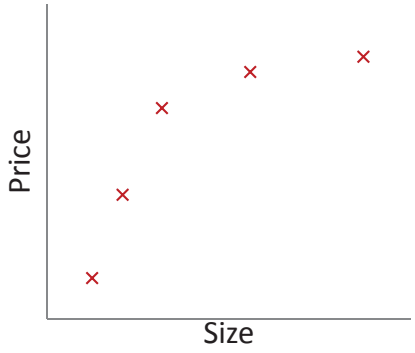
- Note that $\sum_{j=1}^{d}\theta_j^2 = \|\boldsymbol{\theta}_{1:d}\|_2^2$
  - this is magnitude of the feature coefficient vector
- We can also think of this as $\sum_{j=1}^{d}(\theta_j - 0)^2 = \|\boldsymbol{\theta}_{1:d} - \mathbf{0}\|_2^2$
  - L_2-regularization pulls coefficients towards $\mathbf{0}$

Based on slide by Eric Eaton

Understanding Regularization

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

What happens if we set λ to be huge (e.g. 10^{10})?



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Based on slide by Eric Eaton [example by Andrew Ng]

Regularized Linear Regression

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

Gradient Update (single training example)

$$\frac{\partial}{\partial\theta_0}J(\boldsymbol{\theta})$$

$$\frac{\partial}{\partial\theta_j}J(\boldsymbol{\theta})$$

Regularized Linear Regression

$$J(\boldsymbol{\theta}) = \frac{1}{2}\left[(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right]$$

really only $\theta_{1:d}$

Closed-Form Solution

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0$$

Based on slide by Eric Eaton

Probabilistic Interpretation

Why minimize $\sum_{j=1}^{d} \theta_j^2$?

Regularization assumes **prior** on model

- $\boldsymbol{\theta} \sim N(0,\ \lambda^{-1}\boldsymbol{I})$ [multivariate Gaussian]
- $p(\boldsymbol{\theta}) = \dfrac{1}{(2\pi)^{d/2}} \exp\left(-\dfrac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right)$

Maximum *a Posteriori* (MAP) Estimator

Baye's Rule

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \underbrace{p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})}_{\text{posterior}} = \arg\max_{\boldsymbol{\theta}} \underbrace{p(\boldsymbol{y}|\boldsymbol{X}; \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\text{prior}}$$

Try it out

- use log trick
- remember min cost = max probability

Generalizing Regularization

Regularized Linear Regression

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

Generally

$$J_{n,\lambda}(\boldsymbol{\theta}) = \underbrace{R_n(\boldsymbol{\theta})}_{\text{empirical risk}} + \underbrace{\lambda z(\boldsymbol{\theta})}_{\text{regularization}}$$

model fit to data

$$R_n(\boldsymbol{\theta}) = \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$$

$$z(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$$

$z(\boldsymbol{\theta})$ regularization

- biases parameters toward default values (e.g. $\mathbf{0}$)
- resists setting parameters away from zero when data (weakly) tells us otherwise

Common Regularizers

note that sums start from $j = 1$ (do not regularize θ_0)

number of non-zero entries

$$\sum_{j=1}^{d} |\theta_j|^0 = \sum_{j:\theta_j \neq 0} 1$$

$$\|\boldsymbol{\theta}_{[1:d]}\|_0$$

L_0-norm

Minimizing L_0-norm is NP-hard.

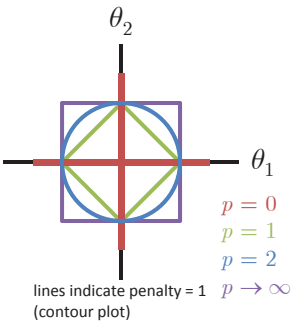
sum of weights

$$\sum_{j=1}^{d} |\theta_j|$$

$$\|\boldsymbol{\theta}_{[1:d]}\|_1$$

L_1-norm

L_1-norm is popular because it results in a sparse solution, but it is not differentiable.



lines indicate penalty = 1 (contour plot)

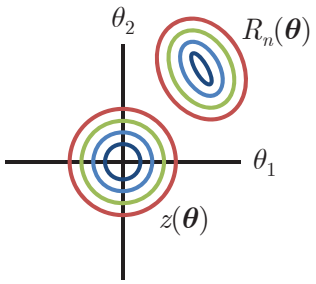
sum of squared weights

$$\sqrt{\sum_{j=1}^{d} |\theta_j|^2}$$

$$\|\boldsymbol{\theta}_{[1:d]}\|_2$$

L_2-norm

L_2-norm is popular because (for some functions) it can be solved directly.



Linear Algebra
What is a norm?
A function that assigns a *strictly positive* length or size to each vector in a vector space, save for the zero vector, which has zero norm.

L_p-norm
- smaller p encourages sparser vectors
- larger p discourages larger weights more

Common Regularizers

Claim: If f and g are convex functions, then so if $f + g$.

Claim: p-norms are convex for $p \geq 1$.

Common names for regularized regression

name	loss	regularization
ordinary least squares (OLS)	squared	none
ridge regression	squared	L_2
lasso regression	squared	L_1
elastic regression	squared	$L_1 + L_2$
logistic regression	logistic	

Summary

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \underbrace{R_n(\boldsymbol{\theta})}_{\text{loss function}} + \underbrace{\lambda z(\boldsymbol{\theta})}_{\text{regularizer}}$$

loss function: negative log likelihood

regularizer: negative log prior

$\boldsymbol{\theta}$ learned (parameter)

λ set by user (hyperparameter)

MLE $\rightarrow$ unregularized

MAP $\rightarrow$ regularized

Be able to answer …

- How do we measure error? (What is the criterion by which we choose $\boldsymbol{\theta}$?)
- What algorithms can we use to optimize this criterion?
  How do these algorithms scale with d and n?
- When $d > n$, there may be directions that remain unconstrained by the data. What do we do?
- How can we constrain $\mathcal{H}$ to achieve better generalization?

Check in

- draw training and test error
- label low / high regularization region
- label overfitting / underfitting region