Lecture 5: Linear Regression with Regularization
CSC 84020 - Machine Learning

Andrew Rosenberg

February 19, 2009

Today

- Linear Regression with Regularization

Recap

Linear Regression

Given a target vector $\mathbf{t}$, and data matrix $\mathbf{X}$.

Goal: Identify the best parameters for a regression function $y = w_0 + \sum_{i=1}^{N} w_i x_i$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

Closed form solution for linear regression

This solution is based on

- Maximum Likelihood estimation under an assumption of Gaussian Likelihood
- Empirical Risk Minimization under an assumption of Squared Error

The extension of Basis Functions gives linear regression significant power.

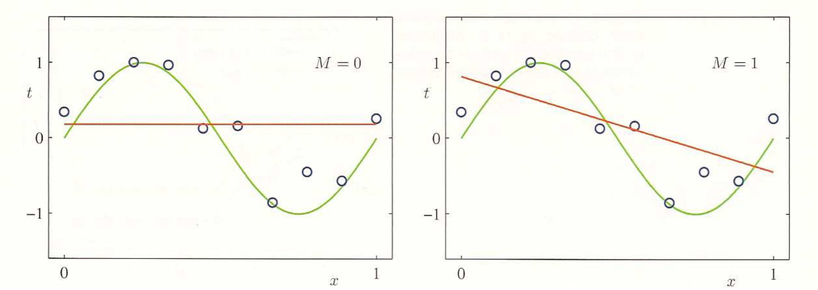
Revisiting overfitting

Overfitting occurs when a model captures idiosyncrasies of the input data, rather than generalizing.

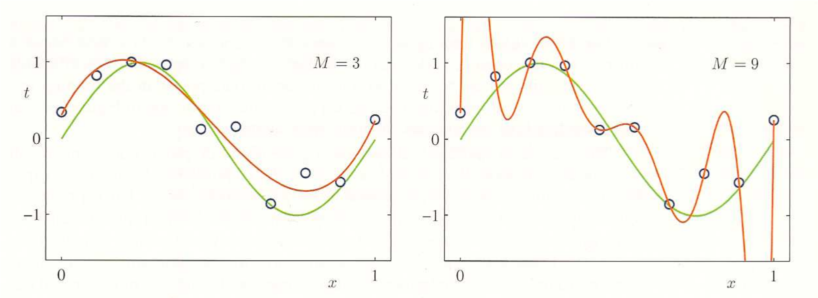
- Too many parameters relative to the amount of training data

For example, an order-N polynomial can be exact fit to $N + 1$ data points.

Overfitting Example



Overfitting Example



Avoiding Overfitting

Ways of detecting/avoiding overfitting.

- Use more data
- Evaluate on a parameter tuning set
- **Regularization**
- Take a Bayesian approach

Regularization

In a Linear Regression model, overfitting is characterized by large parameters.

	$M=0$	$M=1$	$M=3$	$M=9$
w_0	0.19	0.82	0.31	0.35
w_1		-1.27	7.99	232.37
w_2			-25.43	-5321.83
w_3			17.37	48568.31
w_4				-231639.30
w_5				640042.26
w_6				-1061800.52
w_7				1042400.18
w_8				-557682.99
w_9				125201.43

Regularization

Introduce a penalty term for the size of the weights.

Unregularized Regression

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}\{t_n - y(x_n, \mathbf{w})\}^2$$

Regularized Regression
(L2-Regularization or Ridge Regularization)

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Note: Large λ leads to higher complexity penalization.

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = 0$$

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} \sum_{i=0}^{N-1} (y(x_i, \mathbf{w}) - t_i)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right) = 0$$

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right) = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = 0$$

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} \sum_{i=0}^{N-1} (y(x_i, \mathbf{w}) - t_i)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right) = 0$$

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right) = 0$$

$$\nabla_{\mathbf{w}} \left(\frac{1}{2} (\mathbf{t} - \mathbf{X}\mathbf{w})^T (\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \right) = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = 0$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = 0$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{t}$$

Least Squares Regression with L2-Regularization

$$\nabla_{\mathbf{w}} \left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}} \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

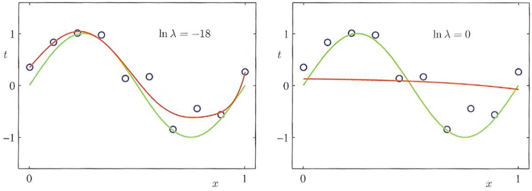
$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = 0$$

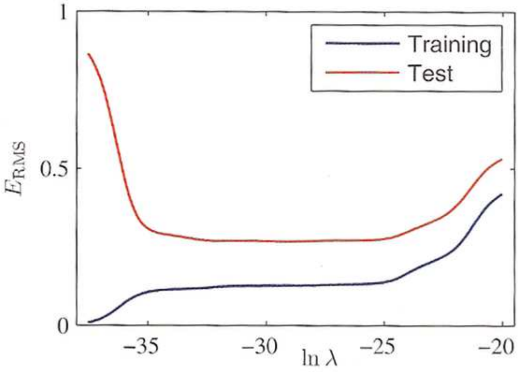
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{t}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}$$

Regularization Results



Regularization Results



Further Regularization

Regularization Approaches

L2-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

L1-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda|\mathbf{w}|_1$$

L0-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda\sum_{n=0}^{N-1}\delta(w_n \neq 0)$$

The **L0-norm** represents the optimal subset of features needed by a Regression model.

Regularization Approaches

L2-Regularization **Closed form** in polynomial time.

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

L1-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda|\mathbf{w}|_1$$

L0-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda\sum_{n=0}^{N-1}\delta(w_n \neq 0)$$

The **L0-norm** represents the optimal subset of features needed by a Regression model.

How can we optimize of these functions?

Further Regularization

Regularization Approaches

L2-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

L1-Regularization Can be **approximated** in poly-time

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda|\mathbf{w}|_1$$

L0-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda\sum_{n=0}^{N-1}\delta(w_n \neq 0)$$

The **L0-norm** represents the optimal subset of features needed by a Regression model.

How can we optimize of these functions?

Regularization Approaches

L2-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

L1-Regularization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda|\mathbf{w}|_1$$

L0-Regularization **NP complete** optimization

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \lambda\sum_{n=0}^{N-1}\delta(w_n \neq 0)$$

The **L0-norm** represents the optimal subset of features needed by a Regression model.

How can we optimize of these functions?

Curse of Dimensionality

Increasing the dimensionality of the feature space exponentially increases the data needs.
Note: The dimensionality of the feature space = The number of features.
What is the message of this?

- Models should be small relative to the amount of available data.
- Dimensionality Reduction techniques – feature selection – can help.
  - L0-regularization is feature selection for linear models.
  - L1- and L2-regularizations approximate feature selection **and** regularize the function.

Curse of Dimensionality Example

Assume a cell requires 100 data points to generalize properly, and 3-ary multinomial features.

- One dimension – requires 300 data points
- Two Dimensions – requires 900 data points
- Three Dimensions – requires 2,700 data points

In this example, for D-dimensional model fitting, the data requirements are $3^D * 10$.

Argument against the **Kitchen Sink** approach.

Bayesians v. Frequentists

What is a Probability?

What is a Probability?

The **Frequentist** position

- A probability is the likelihood that an event will happen.
- It is approximated as the ratio of the number of times the event happened to the total number of events.
- Assessment is very important to select a model.
- Point Estimates are fine $\frac{n}{N}$

What is a Probability?

The **Frequentist** position

- A probability is the likelihood that an event will happen.
- It is approximated as the ratio of the number of times the event happened to the total number of events.
- Assessment is very important to select a model.
- Point Estimates are fine $\frac{n}{N}$

The **Bayesian** position

- A probability is the degree of believability that the event will happen.
- Bayesians require that probabilities be conditioned on data, $p(y|\mathbf{x})$.
- The Bayesian approach "is optimal", given a good model, and good prior and good loss function – don't worry about assessment as much.
- Bayesians say: if you are ever making a point estimate, you've made a mistake. The only valid probabilities are posteriors based on evidence given some prior.

In the previous derivation of the linear regression optimization, we made point estimates for the weight vector, $\mathbf{w}$.

Bayesians would say – "stop right there". Use a distribution over $\mathbf{w}$ to estimate the parameters.

$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

α is a *hyperparameter* over $\mathbf{w}$, where α is the *precision* or inverse variance of the distribution.

So, optimize

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Bayesian Linear Regression

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Again, optimizing the **log** likelihood yields a simpler solution.

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=0}^{N-1} \frac{\beta}{\sqrt{2\pi}} \exp\left\{-\frac{\beta}{2}(t_n - y(x_n, \mathbf{w}))^2\right\}$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2$$

Bayesian Linear Regression

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

Again, optimizing the **log** likelihood yields a simpler solution.

$$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) + \ln p(\mathbf{w}|\alpha)$$

$$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\sum_{n=0}^{N-1}(t_n - y(x_n,\mathbf{w}))^2$$

$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

$$\ln p(\mathbf{w}|\alpha) = \frac{M+1}{2}\ln\alpha - \frac{M+1}{2}\ln 2\pi - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Bayesian Linear Regression

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Again, optimizing the **log** likelihood yields a simpler solution.

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2$$

$$\ln p(\mathbf{w}|\alpha) = \frac{M+1}{2}\ln\alpha - \frac{M+1}{2}\ln 2\pi - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Bayesian Linear Regression

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Again, optimizing the **log** likelihood yields a simpler solution.

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \sum_{n=0}^{N-1} (t_n - y(x_n, \mathbf{w}))^2$$

$$\ln p(\mathbf{w}|\alpha) = \frac{M+1}{2} \ln \alpha - \frac{M+1}{2} \ln 2\pi - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha) = \frac{\beta}{2} \sum_{n=0}^{N-1} (t_n - y(x_n, \mathbf{w}))^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

Broader Context

Overfitting is bad.

Bayesians v. Frequentists.

Does it matter which camp you lie in?

Not particularly, but Bayesian approaches allow us some useful interesting and principled tools.

Bye

- Next
  - Categorization
    - Logistic Regression
    - Naive Bayes