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

$$y = w_0 + \sum_{i=1}^{N} w_i x_i$$

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

\[
\begin{align*}
M = 0 & : \\
M = 1 & :
\end{align*}
\]
Overfitting Example

\[ M = 3 \]

\[ M = 9 \]
Ways of detecting/avoiding overfitting.

- Use more data
- Evaluate on a parameter tuning set
- **Regularization**
- Take a Bayesian approach
In a Linear Regression model, overfitting is characterized by large parameters.

<table>
<thead>
<tr>
<th></th>
<th>$M = 0$</th>
<th>$M = 1$</th>
<th>$M = 3$</th>
<th>$M = 9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_0$</td>
<td>0.19</td>
<td>0.82</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>$w_1$</td>
<td>-1.27</td>
<td>7.99</td>
<td>-25.43</td>
<td>232.37</td>
</tr>
<tr>
<td>$w_2$</td>
<td>17.37</td>
<td>48568.31</td>
<td>-5321.83</td>
<td></td>
</tr>
<tr>
<td>$w_3$</td>
<td>-231639.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_4$</td>
<td>640042.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_5$</td>
<td>-1061800.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_6$</td>
<td>1042400.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_7$</td>
<td>-557682.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_8$</td>
<td>125201.43</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Introduce a penalty term for the size of the weights.

**Unregularized Regression**

\[
E(w) = \frac{1}{2} \sum_{n=0}^{N-1} \left( t_n - y(x_n, w) \right)^2
\]

**Regularized Regression**

(L2-Regularization or Ridge Regularization)

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

Note: Large $\lambda$ leads to higher complexity penalization.
Least Squares Regression with L2-Regularization

$$\nabla_w E(w) = 0$$
Least Squares Regression with L2-Regularization

\[ \nabla_w (E(w)) = 0 \]

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

\[ \nabla_w \left( \frac{1}{2} \|t - Xw\|^2 + \frac{\lambda}{2} \|w\|^2 \right) = 0 \]
Least Squares Regression with L2-Regularization

\[ \nabla_w(E(w)) = 0 \]

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

\[ \nabla_w \left( \frac{1}{2} \| t - Xw \|^2 + \frac{\lambda}{2} \| w \|^2 \right) = 0 \]

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0 \]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \lambda w = 0 \]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \lambda w = 0 \]

\[ -X^T t + X^T Xw + \lambda lw = 0 \]
Least Squares Regression with L2-Regularization

\[
\nabla_w \left( \frac{1}{2} (t - Xw)^T(t - Xw) + \frac{\lambda}{2} w^T w \right) = 0
\]

\[
-X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0
\]

\[
-X^T t + X^T Xw + \lambda w = 0
\]

\[
-X^T t + X^T Xw + \lambda Iw = 0
\]

\[
-X^T t + (X^T X + \lambda I)w = 0
\]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0 \]

\[ -X^T t + X^T Xw + \lambda w = 0 \]

\[ -X^T t + X^T Xw + \lambda I w = 0 \]

\[ -X^T t + (X^T X + \lambda I)w = 0 \]

\[ (X^T X + \lambda I)w = X^T t \]
Least Squares Regression with L2-Regularization

\[ \nabla_w \left( \frac{1}{2} (t - Xw)^T (t - Xw) + \frac{\lambda}{2} w^T w \right) = 0 \]

\[-X^T t + X^T Xw + \nabla_w \left( \frac{\lambda}{2} w^T w \right) = 0 \]

\[-X^T t + X^T Xw + \lambda w = 0 \]

\[-X^T t + X^T Xw + \lambda Iw = 0 \]

\[-X^T t + (X^T X + \lambda I)w = 0 \]

\[(X^T X + \lambda I)w = X^T t \]

\[w = (X^T X + \lambda I)^{-1} X^T t \]
Regularization Results
Regularization Results

![Graph showing regularization results with two curves: Training and Test. The y-axis represents $E_{RMS}$ ranging from 0 to 1, and the x-axis represents $\ln \lambda$ ranging from -35 to -20.](image)
Further Regularization

Regularization Approaches

L2-Regularization

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

L1-Regularization

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

L0-Regularization

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

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

Regularization Approaches

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

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

L1-Regularization

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

L0-Regularization

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

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

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

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

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

L0-Regularization

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

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

How can we optimize these functions?
Further Regularization

Regularization Approaches

L2-Regularization

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

L1-Regularization

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

L0-Regularization \textbf{NP complete} optimization

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

The \textbf{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.
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 $3^D \times 10$.

Argument against the **Kitchen Sink** approach.
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 $\frac{n}{N}$
Bayesians v. Frequentists

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, $p(y|x)$.
- 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.

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

\( \alpha \) is a hyperparameter over \( \mathbf{w} \), where \( \alpha \) 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

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]

Again, optimizing the \textbf{log} likelihood yields a simpler solution.

\[
\ln p(t|x, w, \beta) + \ln p(w|\alpha)
\]

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

\[
\ln p(t|x, w, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \sum_{n=0}^{N-1} (t_n - y(x_n, w))^2
\]
Bayesian Linear Regression

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]

Again, optimizing the log likelihood yields a simpler solution.

\[ \ln p(t|x, w, \beta) + \ln p(w|\alpha) \]

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

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

\[ \ln p(w|\alpha) = \frac{M + 1}{2} \ln \alpha - \frac{M + 1}{2} \ln 2\pi - \frac{\alpha}{2} w^T w \]
Bayesian Linear Regression

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]

Again, optimizing the log likelihood yields a simpler solution.

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

\[
\ln p(w|\alpha) = \frac{M + 1}{2} \ln \alpha - \frac{M + 1}{2} \ln 2\pi - \frac{\alpha}{2} w^T w
\]
Bayesian Linear Regression

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]

Again, optimizing the \textbf{log} likelihood yields a simpler solution.

\[
\ln p(t|x, w, \beta) + \ln p(w|\alpha)
\]

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

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

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

Categorization

- Logistic Regression
- Naive Bayes