Machine learning - HT 2016
4. Basis Expansion, Regularization, Validation

Varun Kanade

University of Oxford
February 03, 2016
Introduce **basis function** to go beyond linear regression

Understanding the tradeoff between **bias** and **variance**

**Overfitting**: What happens when we make models too complex

**Regularization** as a means to control model complexity

**Cross-validation** to perform model selection
Basis Expansion
Basis Expansion
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, x^4] \]

\[ w_1 + w_2x + w_3x^2 + w_4x^3 + w_4x^5 = \phi(x) \cdot [w_1, w_2, w_3, w_4, w_5] \]
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, x^4] \]

\[ w_1 + w_2x + w_3x^2 + w_4x^3 + w_4x^5 = \phi(x) \cdot [w_1, w_2, w_3, w_4, w_5] \]
Basis Expansion

\[ \phi(x) = [1, x, x^2] \]
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, \ldots, x^9] \]
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, \ldots, x^d] \]

How can we avoid overfitting?
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, \ldots, x^d] \]

How can we avoid overfitting?

Does more data help?
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, \ldots, x^d] \]

How can we avoid overfitting?

Does more data help?
Basis Expansion

\[ \phi(x) = [1, x, x^2, x^3, \ldots, x^d] \]

How can we avoid overfitting?

Does more data help?
Bias Variance Tradeoff

- For linear model, more data would make little difference

- **Bias** results from model being simpler than the “truth”
- High bias results in underfitting
What happens when we fit model on different (randomly drawn) training datasets?

- **Variance** arises when the (complex) model is sensitive to fluctuations in training dataset
- **Variance** results in overfitting
Bias Variance Tradeoff

- When does more data help?
- Error = $\text{Bias}^2 + \text{Variance} + \text{Noise}$ (Exercise for linear regression)

For more complex models, difficult to visually overfitting and underfitting

- Keep aside some points as “test set”
Learning Curves

- Suppose we have a training set and test set
- Train on increasing sizes of the training set, and plot the errors

Once training and test error approach each other, then more data won’t help!
Basis Expansion Using Kernels

We can use kernels as features, e.g., radial basis functions (RBFs)

Feature expansion:

$$\phi(x) = [1, \kappa(x, \mu_1, \sigma), \ldots, \kappa(x, \mu_d, \sigma)]$$

Model: $$y = \phi(x)^T w + \text{noise}$$

$$\kappa(x, \mu_i, \sigma) = e^{-\frac{||x-\mu_i||^2}{2\sigma^2}}$$

\[ y = 0 + 0.5 \kappa(x, \mu_1, \sigma^2) + 0.7 \kappa(x, \mu_2, \sigma^2) + 1.1 \kappa(x, \mu_3, \sigma^2) \]
Basis Expansion Using Kernels

As in the case of polynomials, the width $\sigma$ can cause overfitting or underfitting.

Image Source: K. Murphy (2012)
Polynomial Basis Expansion in Higher Dimensions

We are basically fitting linear models (at the cost of increasing dimensions)

\[ y = \phi(x) + \text{noise} \]

Linear Model: \( \phi(x) = [1, x_1, x_2] \)

Quadratic Model: \( \phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2] \)

How many dimensions do you get for degree \( d \) polynomials over \( n \) variables? grows as \( n^d \) \(
\text{if } n = 1000, \; d = 20, \; \text{this is } 10^{60!} \)
Overfitting!

In high dimensions, we can have many many parameters!

With 100 variables and degree 10 we have $\sim 10^{20}$ parameters!

Enrico Fermi to Freeman Dyson

“I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.” [video]

How do we prevent overfitting?
How does overfitting occur?

Suppose $X_{100 \times 100}$ with every entry $\mathcal{N}(0, 1)$

And let $y_i = x_{i,1} + \mathcal{N}(0, \sigma^2)$, for $\sigma = 0.2$
Ridge Regression

Say our data is \( \{(x_i, y_i)\}_{i=1}^m \), where \( x \in \mathbb{R}^N \) where \( N \) is really really large!

We used the squared loss

\[
L(w) = (Xw - y)^T (Xw - y)
\]

and obtained the estimate

\[
w = (X^T X)^{-1} X^T y
\]

Suppose we want \( w \) to be “small” (\textit{weight decay})

\[
L(w) = (Xw - y)^T (Xw - y) + \lambda w^T w - \lambda w_1^2
\]

We will not regularize the “bias” (or constant) term

\textbf{Exercise:} If all \( x_i \) (except \( x_1 \)) have mean 0 and \( y \) has mean 0, \( w_1 = 0 \)
Deriving Estimate for Ridge Regression

\[ L(w) = (Xw - y)^T (Xw - y) + \lambda w^T w \]

\[ \nabla_w L = 2X^T x w - 2x^T y + 2\lambda w = 0 \]

\[ \Rightarrow w = (X^T x + \lambda I)^{-1} x^T y \]

*Positive semi-definite*

\[ \Rightarrow \exists \gamma \neq 0, \text{ s.t. } (X^T x + \lambda I) \gamma = 0 \]

\[ \Rightarrow X^T x + \lambda I \text{ is non-singular} \]
Ridge Regression

Objective/Loss: \( L(w) = (Xw - y)^T(Xw - y) + \lambda w^T w \)

Estimate: \( \hat{w} = (X^TX + \lambda I)^{-1}X^Ty \)

Estimate depends on the scaling of the inputs

Common practice: Normalise all input dimensions

Ridge penalizes all weights equally.

Good idea to make all features have 0 mean and variance 1.
Ridge Regression and MAP Estimation

Objective/Loss: \( L(w) = (Xw - y)^T(Xw - y) + \lambda w^T w \)

\[
\exp \left( \frac{-1}{2\sigma^2} L(w) \right) = \exp \left( -\frac{1}{2} (y - Xw)^T \Sigma^{-1} (y - Xw) \right) \cdot \exp \left( -\frac{1}{2} w^T \Lambda^{-1} w \right)
\]

\[
\Sigma = \begin{pmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2 \\
\end{pmatrix}
\]

\[
\Lambda = \begin{pmatrix}
\frac{\sigma^2}{\phi^2} & 0 & \cdots & 0 \\
0 & \frac{\sigma^2}{\phi^2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{\sigma^2}{\phi^2} \\
\end{pmatrix}
\]

\[
= C \cdot \mathcal{N}(y | Xw, \Sigma) \cdot \mathcal{N}(w | 0, \Lambda)
\]

normalization

likelihood of \( y | Xw \)

prior on \( w \)
Bayesian View of Machine Learning

General Formulation

Prior \( p(w) \) on \( w \)

Model \( p(y \mid x, w) \)

Linear Regression

\[
p(w) = \mathcal{N}(w \mid 0, \tau^2 I_n)
\]

\[
p(y \mid x, w) = \mathcal{N}(y \mid x^T w, \sigma^2)
\]

Compute posterior given data \( D = \{(x_i, y_i)\}_{i=1}^m \)

\[
p(w \mid D) \propto p(D \mid w) \cdot p(w)
\]

Bayes Rule:

\[
p(A \mid B) = \frac{p(B \mid A) p(A)}{p(B)}
\]

The normalization \( p(D) = \int p(D \mid w) p(w) \) can be difficult to compute in general. However, not needed for MAP estimate.

Mode of posterior distribution.

Maximum A Posteriori (MAP) estimate.

Ridge Regression = MAP estimation with Gaussian priors.
Bayesian View of Regression

Making a prediction on a new point $x_{\text{new}}$

$$p(y \mid D, x_{\text{new}}) = \int_{w} p(y \mid w, x_{\text{new}})p(w \mid D)dw$$

For linear regression

$$p(y \mid D, x_{\text{new}}) = \int_{w} \mathcal{N}(y \mid x_{\text{new}}^Tw, \sigma^2) \cdot \mathcal{N}(w \mid w_{\text{map}}, (X^TX + \frac{\sigma^2}{\tau^2}I)^{-1}) dw$$

$$= \mathcal{N}(y \mid x_{\text{new}}^Tw_{\text{map}}, \sigma^2 (1 + x_{\text{new}}^T(X^TX + \frac{\sigma^2}{\tau^2}I)^{-1}x_{\text{new}}$$

(For details see Murphy Sec. 7.6.1 & 7.6.2)
Prediction MAP vs Fully Bayesian

**MAP Approach**

\[ y_{\text{new}} \sim \mathcal{N}(x_{\text{new}}^T w_{\text{map}}, \sigma^2) \]

**Bayesian Approach**

\[ y_{\text{new}} \sim \mathcal{N}(x_{\text{new}}^T w_{\text{map}}, \sigma_X^2) \]

\[ \sigma_X^2 = \sigma^2 \left(1 + x_{\text{new}}^T (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} x_{\text{new}}\right)^{-1} \]
Ridge Regression

Minimize:

$$(Xw - y)^T(Xw - y) + \lambda w^T w$$

such that $w^T w \leq R$

Minimize $(Xw - y)^T(Xw - y)$
Ridge Regression

Image Source: Hastie, Tibshirani, Friedman (2013)
Lasso Regression

Minimize:

$$(Xw - y)^T(Xw - y) + \lambda \sum_{i=1}^{N} |w_i|$$

such that $\sum_{i=1}^{N} |w_i| \leq R$

No closed form solution.
Needs convex optimization methods.

Solutions are likely to be genuinely sparse if using $L_1$-regularization.
Lasso Regression

Image Source: Hastie, Tibshirani, Friedman (2013)
Regularization: Ridge or Lasso

(Example from Slide 12)

No regularization

Ridge

Lasso

Very few non-zero features in Lasso.
Choosing Parameters

Before we were just trying to find $w$

Now, we have to worry about how to choose $\lambda$ for ridge or Lasso

If we use kernels, we also have to pick the width $\sigma$

If we use higher degree polynomials, we have to pick $d$
Cross Validation

- Keep a part of data as “validation” or “test” set
- Look for error on “training” and “validation” sets

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Train Error(%)</th>
<th>Test Error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0</td>
<td>89</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>27</td>
</tr>
</tbody>
</table>
$k$-Fold Cross Validation

What do we do when data is scarce?

- Divide data into $k$ parts
- Use $k - 1$ parts for training and 1 part as validation
- When $k = m$ (the number of datapoints), we get LOOCV (Leave one out cross validation)
Suppose you do all the right things

- Train on the training set
- Choose hyperparameters using proper validation
- Test on the test set, and your error is high!

What would you do?
Winning Kaggle without reading the data!

Suppose the task is to predict $m$ binary labels

Algorithm (Wacky Boosting):

1. Choose $y^1, \ldots, y^k \in \{0, 1\}^m$ uniformly
2. Set $I = \{i \mid \text{accuracy}(y^i) > 51\%\}$
3. Output $\hat{y}_j = \text{majority}\{y^i_j \mid i \in I\}$

Source: blog.mrtz.org