CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019
• Lab 2 due **THURSDAY** at midnight
• Lab 3 released today, due next Thursday
• Scribe notes for **extra credit!**
• Let me know of any partner issues (anytime)
• Reading posted (logistic regression & naïve bayes

---

**Research Talk**

**Thursday, February 14, 2019**

11:30-12:30 p.m.

Science Center 256

**Anne Cocos**

Department of Computer and Information Science

University of Pennsylvania

“Semantic Structure from Paraphrase Pairs”
Outline for February 13

• Polynomial regression

• Regularization

• Linear regression for classification

• Begin: logistic regression
Outline for February 13

• Polynomial regression

• Regularization

• Linear regression for classification

• Begin: logistic regression
Polynomial Regression

• Can be thought of as regular linear regression with a change of basis
Polynomial Regression

before: \( h_b^*(\tilde{x}) = \tilde{b}^T \tilde{x} \)

\[ \tilde{x} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_p \end{bmatrix} \]

find coefficient for each term

quadratic model

Generalized linear model

\( h_b^*(\tilde{x}) = \tilde{b}^T \Phi(\tilde{x}) \)

some transformation

\[ \tilde{b} = (\Phi(\tilde{x})) \phi \]

\( n \) data points

\( \Rightarrow \) fit

\[ n=3, \quad d=2 \]
Analytic Solution

\[ \hat{b} = (\Phi^T \Phi)^{-1} (\Phi^T \vec{y}) \]

n data points, degree \( d = n - 1 \)

\( \Rightarrow \) fit perfectly!

\[
\begin{align*}
\text{n=3, } d=2 \\
\text{n=4, } d=3
\end{align*}
\]
\[ P = 1, \quad d = \text{degree} \]

\[ h_b(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_d x^d \]

In general, \( p \) features, degree \( d \) \( \Rightarrow \) \( pd + 1 \) coeffs.

\( \phi(x) = \begin{bmatrix} 1 & x & x^2 & \ldots & x^p \end{bmatrix} \)

\[ b^T \phi(x) \]

\[ SGR \]

\[ b_0 \leftarrow b_i \]

Hyperparameters
Outline for February 13

• Polynomial regression

• Regularization

• Linear regression for classification

• Begin: logistic regression
Generalization error

- Example: price vs. size (i.e. of a house or car)
Generalization error

- Example: price vs. size (i.e. of a house or car)

- Underfitting (high bias)
- Correct fit
- Overfitting (high variance)

\[ \text{Price} \]

\[ b_0 + b_1 x \]

\[ b_0 + b_1 x + b_2 x^2 \]

\[ b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 \]

Adapted from slide by Jessica Wu
[example by Andrew Ng]
Generalization error

- Example: price vs. size (i.e. of a house or car)

**Structural error:**
Hypothesis space cannot model true relationship

⇒ More data doesn’t help
⇒ Need a more flexible model

**Estimation (approximation) error:**
Hypothesis space *can* model true relationship, BUT hard to identify correct model due to large hypothesis space, small $n$, or noise
⇒ Reduce hypothesis space
⇒ Add more data
Regularization

What if ...

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

What if ...

• we have a limited # of training examples \((n<p)\), or
• we want to automatically control the complexity of the learned hypothesis?

Idea: penalize large values of \(b_j\)

Why prefer small weights?

• if large weights, small change in feature can result in large change in prediction
• prevent giving too much weight to any one feature
• might prefer zero weight for useless features
Common Regularizers

$L_0$ norm

- Number of non-zero entries
- Minimizing $L_0$ norm is NP hard
Common Regularizers

\[ \|b\|_0 = \sum_{j: b_j \neq 0} 1 \]

**L₀ norm**
- Number of non-zero entries
- Minimizing L₀ norm is NP hard

\[ \|b\|_1 = \sum_{j=1}^{p} |b_j| \]

**L₁ norm**
- Sum of magnitude of weights
- Not differentiable
Common Regularizers

\[ \|b\|_0 = \sum_{j : b_j \neq 0} 1 \]
\[ \|b\|_1 = \sum_{j=1}^{p} |b_j| \]
\[ \|b\|_2 = \sqrt{\sum_{j=1}^{p} b_j^2} \]

**L\(_0\) norm**
- Number of non-zero entries
- Minimizing \(L_0\) norm is NP hard

**L\(_1\) norm**
- Sum of magnitude of weights
- Not differentiable

**L\(_2\) norm**
- Sum of squared weights
- Differentiable

Slide adapted from Jessica Wu
Common Regularizers

$L_0$ norm

- Number of non-zero entries
- Minimizing $L_0$ norm is NP hard

$L_1$ norm

- Sum of magnitude of weights
- Not differentiable

$L_2$ norm

- Sum of squared weights
- Differentiable

\[
\|b\|_0 = \sum_{j: b_j \neq 0} 1
\]

\[
\|b\|_1 = \sum_{j=1}^{p} |b_j|
\]

\[
\|b\|_2 = \sqrt{\sum_{j=1}^{p} b_j^2}
\]

lines indicate penalty = 1
(contour plot)
Regularization

\[ J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \frac{\lambda}{2} \sum_{j=1}^{p} (\theta_j)^2 \]

\( \lambda \geq 0 \) is the regularization parameter.
- Small \( \lambda \): more fit to training data.
- Large \( \lambda \): keeps weights small and increases more generalization.

Choose \( \lambda = 10^0 \) (very large)

Do not regularize \( \theta_0 \)

\[ \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \]
SGD

\[ b_0 \leftarrow b_0 - \alpha (b^T x_i - y_i) \]

\[ b_j \leftarrow b_j - \alpha \left[ (b^T x_i - y_i) x_{ij} + \lambda b_j \right] \]

Hyperparameters

\[ \gamma = (1 - \alpha \lambda) b_j - \alpha \left[ (b^T x_i - y_i) x_{ij} \right] \]

Pulling terms towards 0 each iteration.

Analytic

\[ J(b) = (Xb - y)^T (Xb - y) + \lambda b^T b \]

\[ \hat{b} = (X^T X + \lambda I)^{-1} X^T y \]

\[ I = \begin{bmatrix} 0 & 0 & 0 & 6 \\
0 & 0 & 0 & 6 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \end{bmatrix} \]
Outline for February 13

• Polynomial regression

• Regularization

• Linear regression for classification

• Begin: logistic regression
\[ \hat{y} = 1 \text{ if } \bar{b} \frac{T}{x} \geq 0.5 \]
\[ \hat{y} = 0 \text{ o.w.} \]