Support Vector Machine I

Jia-Bin Huang
Virginia Tech
Administrative

• Please use piazza. No emails.

• HW 0 grades are back. Re-grade request for one week.

• HW 1 due soon.

• HW 2 on the way.
Regularized logistic regression

\[ J(\theta) = \frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log \left(1 - h_\theta(x^{(i)})\right) + \frac{\lambda}{2} \sum_{j=1}^{n} \theta_j^2 \right] \]

- Cost function:

\[ h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1^3 x_2 + \theta_7 x_1 x_2^3 + \ldots) \]
Gradient descent (Regularized)

Repeat 

\[
\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})
\]

\[
h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}
\]

\[
\theta_j := \theta_j - \alpha \left[ \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \lambda \theta_j \right]
\]

\[
\frac{\partial}{\partial \theta_j} J(\theta)
\]

Slide credit: Andrew Ng
$|\theta|_1$: Lasso regularization

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

LASSO: Least Absolute Shrinkage and Selection Operator

Regression shrinkage and selection via the lasso
We propose a new method for estimation in linear models. The lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ...
Single predictor: Soft Thresholding

- \( \text{minimize}_\theta \frac{1}{2m} \sum_{i=1}^{m} (x^{(i)} \theta - y^{(i)})^2 + \lambda |\theta|_1 \)

- \( \hat{\theta} = \begin{cases} 
\frac{1}{m} < x, y > -\lambda & \text{if } \frac{1}{m} < x, y > > \lambda \\
0 & \text{if } \frac{1}{m} | < x, y > | \leq \lambda \\
\frac{1}{m} < x, y > +\lambda & \text{if } \frac{1}{m} < x, y > < -\lambda 
\end{cases} \)

\[
\hat{\theta} = S_\lambda \left( \frac{1}{m} < x, y > \right)
\]

Soft Thresholding operator \( S_\lambda (x) = \text{sign}(x)(|x| - \lambda)_+ \)
Multiple predictors: Cyclic Coordinate Descenet

- minimize $\theta \frac{1}{2m} \sum_{i=1}^{m} \left( x_j^{(i)} \theta_j + \sum_{k \neq j} x_{ij}^{(i)} \theta_k - y^{(i)} \right)^2 + \lambda \sum_{k \neq j} |\theta_k| + \lambda |\theta_j|_1$

For each $j$, update $\theta_j$ with

$$\text{minimize}_\theta \frac{1}{2m} \sum_{i=1}^{m} \left( x_j^{(i)} \theta_j - r_j^{(i)} \right)^2 + \lambda |\theta_j|_1$$

where $r_j^{(i)} = y^{(i)} - \sum_{k \neq j} x_{ij}^{(i)} \theta_k$
L1 and L2 balls

## Terminology

### Regularization function

<table>
<thead>
<tr>
<th>Regularization function</th>
<th>Name</th>
<th>Solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|\theta|<em>2^2 = \sum</em>{j=1}^{n} \theta_j^2$</td>
<td>Tikhonov regularization, Ridge regression</td>
<td>Close form</td>
</tr>
<tr>
<td>$|\theta|<em>1 = \sum</em>{j=1}^{n}</td>
<td>\theta_j</td>
<td>$</td>
</tr>
<tr>
<td>$\alpha</td>
<td>|\theta|_1 + (1 - \alpha) |\theta|_2^2$</td>
<td>Elastic net regularization</td>
</tr>
</tbody>
</table>
Things to remember

- **Overfitting**
  - Complex model: doing well on the training set, but perform poorly on the testing set

- **Cost function**
  - Norm penalty: L1 or L2

- **Regularized linear regression**
  - Gradient decent with weight decay (L2 norm)

- **Regularized logistic regression**
  - Gradient decent with weight decay (L2 norm)
Support Vector Machine

• Cost function

• Large margin classification

• Kernels

• Using an SVM
Support Vector Machine

- Cost function
- Large margin classification
- Kernels
- Using an SVM
Logistic regression

\[ h_\theta(x) = g(\theta^\top x) \]

\[ g(z) = \frac{1}{1 + e^{-z}} \]

Suppose predict “y = 1” if \( h_\theta(x) \geq 0.5 \)

predict “y = 0” if \( h_\theta(x) < 0.5 \)

\[ z = \theta^\top x \geq 0 \]

\[ z = \theta^\top x < 0 \]

Slide credit: Andrew Ng
Alternative view

\[ h_\theta(x) = g(\theta^\top x) \]
\[ g(z) = \frac{1}{1 + e^{-z}} \]

If “\( y = 1 \)”, we want \( h_\theta(x) \approx 1 \)

If “\( y = 0 \)”, we want \( h_\theta(x) \approx 0 \)
Cost function for **Logistic Regression**

\[
\text{Cost}(h_\theta(x), y) = \begin{cases} 
-\log(h_\theta(x)) & \text{if } y = 1 \\
-\log(1 - h_\theta(x)) & \text{if } y = 0 
\end{cases}
\]

if \( y = 1 \)

if \( y = 0 \)

Slide credit: Andrew Ng
Alternative view of logistic regression

• Cost($h_\theta(x), y$) = $-y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x))$
  
  $= y \left(-\log\left(\frac{1}{1 + e^{-\theta^\top x}}\right)\right) + (1 - y) \left(-\log\left(1 - \frac{1}{1 + e^{-\theta^\top x}}\right)\right)$

if $y = 1$

if $y = 0$
Logistic regression (logistic loss)

$$
\min_{\theta} \frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \left( -\log \left( h_{\theta}(x^{(i)}) \right) \right) + (1 - y^{(i)}) \left( -\log \left( 1 - h_{\theta}(x^{(i)}) \right) \right) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_{j}^2
$$

Support vector machine (hinge loss)

$$
\min_{\theta} \frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \cos1(\theta^T x^{(i)}) + (1 - y^{(i)}) \cos0(\theta^T x^{(i)}) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_{j}^2
$$

if $y = 1$

$$
-\log \left( \frac{1}{1 + e^{-\theta^T x}} \right)
$$

if $y = 0$

$$
-\log \left( 1 - \frac{1}{1 + e^{-\theta^T x}} \right)
$$
Optimization objective for SVM

$$\min_{\theta} \frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \cos_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \cos_0(\theta^T x^{(i)}) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$

1) Multiply $\frac{1}{m}$
2) Multiply $C = \frac{1}{\lambda}$

$$\min_{\theta} C \left[ \sum_{i=1}^{m} y^{(i)} \cos_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \cos_0(\theta^T x^{(i)}) \right] + \sum_{j=1}^{n} \theta_j^2$$

Slide credit: Andrew Ng
Hypothesis of SVM

$$
\min_\theta C \left[ \sum_{i=1}^{m} y^{(i)} \text{cost}_1 (\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0 (\theta^T x^{(i)}) \right] + \sum_{j=1}^{n} \theta_j^2
$$

• Hypothesis

$$
h_\theta (x) = \begin{cases} 
1 & \text{if } \theta^T x \geq 0 \\
0 & \text{if } \theta^T x < 0 
\end{cases}
$$
Support Vector Machine

• Cost function

• Large margin classification

• Kernels

• Using an SVM
Support vector machine

\[
\min_{\theta} C \left[ \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) \right] + \sum_{j=1}^{n} \theta_j^2
\]

If \( y = 1 \), we want \( \theta^T x \geq 1 \) (not just \( \geq 0 \))
If \( y = 0 \), we want \( \theta^T x \leq -1 \) (not just \( < 0 \))
SVM decision boundary

\[
\min_{\theta} C \left[ \sum_{i=1}^{m} y^{(i)} \cos_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \cos_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2
\]

• Let’s say we have a very large \( C \) ...

• Whenever \( y^{(i)} = 1 \):
  \( \theta^T x^{(i)} \geq 1 \)

• Whenever \( y^{(i)} = 0 \):
  \( \theta^T x^{(i)} \leq -1 \)

\[
\min_{\theta} \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 \\
\text{s.t.} \quad \theta^T x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1 \\
\quad \theta^T x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0
\]

Slide credit: Andrew Ng
SVM decision boundary: Linearly separable case
SVM decision boundary: Linearly separable case

Slide credit: Andrew Ng
Large margin classifier in the presence of outlier

\[ x_2 \]

\[ x_1 \]

\( C \) very large

\( C \) not too large

Slide credit: Andrew Ng
Vector inner product

\[ u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \]

\[ ||u|| = \text{length of vector } u \]

\[ = \sqrt{u_1^2 + u_2^2} \in \mathbb{R} \]

\[ p = \text{length of projection of } v \text{ onto } u \]

\[ u^\top v = p \cdot ||u|| = u_1 v_1 + u_2 v_2 \]

Slide credit: Andrew Ng
SVM decision boundary

\[
\min_{\theta} \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 \\
\text{s. t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1 \\
\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0 \\
\text{Simplication:} \quad \theta_0 = 0, n = 2
\]

What’s \( \theta^\top x^{(i)} \)?

\[
\theta^\top x^{(i)} = p(i) \|\theta\|^2
\]

\[
\frac{1}{2} \sum_{j=1}^{n} \theta_j^2 = \frac{1}{2} (\theta_1^2 + \theta_2^2) = \frac{1}{2} \left( \sqrt{\theta_1^2 + \theta_2^2} \right)^2 = \frac{1}{2} \|\theta\|^2
\]
SVM decision boundary

\[
\min_{\theta} \frac{1}{2} \|\theta\|^2
\]

s.t.

\[
p^{(i)} \|\theta\|^2 \geq 1 \quad \text{if } y^{(i)} = 1
\]

\[
p^{(i)} \|\theta\|^2 \leq -1 \quad \text{if } y^{(i)} = 0
\]

Simplication: \(\theta_0 = 0, n = 2\)

\[
p^{(1)}, p^{(2)} \text{ small } \rightarrow \|\theta\|^2 \text{ large}
\]

\[
p^{(1)}, p^{(2)} \text{ large } \rightarrow \|\theta\|^2 \text{ can be small}
\]
Support Vector Machine

- Cost function
- Large margin classification
- Kernels
- Using an SVM
Non-linear decision boundary

Predict $y = 1$ if

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \geq 0$$

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \cdots$$

$$f_1 = x_1, f_2 = x_2, f_3 = x_1 x_2, \cdots$$

Is there a different/better choice of the features $f_1, f_2, f_3, \cdots$?

Slide credit: Andrew Ng
Give $x$, compute new features depending on proximity to landmarks $l^{(1)}$, $l^{(2)}$, $l^{(3)}$

\[
\begin{align*}
    f_1 &= \text{similarity}(x, l^{(1)}) \\
    f_2 &= \text{similarity}(x, l^{(2)}) \\
    f_3 &= \text{similarity}(x, l^{(3)})
\end{align*}
\]

Gaussian kernel

\[
\text{similarity}(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)
\]
Predict $y = 1$ if

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$$

Ex: $\theta_0 = -0.5, \theta_1 = 1, \theta_2 = 1, \theta_3 = 0$

$f_1 = \text{similarity}(x, l^{(1)})$
$f_2 = \text{similarity}(x, l^{(2)})$
$f_3 = \text{similarity}(x, l^{(3)})$
Choosing the landmarks

• Given $x$

$$f_i = \text{similarity}(x, l^{(i)}) = \exp\left(-\frac{||x - l^{(i)}||^2}{2\sigma^2}\right)$$

Predict $y = 1$ if $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

Where to get $l^{(1)}, l^{(2)}, l^{(3)}, \ldots$?
SVM with kernels

- Given \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\)
- Choose \(l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, l^{(3)} = x^{(3)}, \ldots, l^{(m)} = x^{(m)}\)

- Given example \(x\):
  - \(f_1 = \text{similarity}(x, l^{(1)})\)
  - \(f_2 = \text{similarity}(x, l^{(2)})\)
  - \(\ldots\)

- For training example \((x^{(i)}, y^{(i)})\):
  - \(x^{(i)} \rightarrow f^{(i)}\)

\[
 f = \begin{bmatrix}
 f_0 \\
 f_1 \\
 f_2 \\
 \vdots \\
 f_m 
\end{bmatrix}
\]
SVM with kernels

• Hypothesis: Given $x$, compute features $f \in \mathbb{R}^{m+1}$
  • Predict $y = 1$ if $\theta^T f \geq 0$

• Training (original)

$$
\min_{\theta} C \left[ \sum_{i=1}^{m} y^{(i)} \cos_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \cos_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2
$$

• Training (with kernel)

$$
\min_{\theta} C \left[ \sum_{i=1}^{m} y^{(i)} \cos_1(\theta^T f^{(i)}) + (1 - y^{(i)}) \cos_0(\theta^T f^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{m} \theta_j^2
$$
SVM parameters

• $C \left(= \frac{1}{\lambda}\right)$
  - Large $C$: Lower bias, high variance.
  - Small $C$: Higher bias, low variance.

• $\sigma^2$
  - Large $\sigma^2$: features $f_i$ vary more smoothly.
    • Higher bias, lower variance
  - Small $\sigma^2$: features $f_i$ vary less smoothly.
    • Lower bias, higher variance

Slide credit: Andrew Ng
SVM song

• https://www.youtube.com/watch?v=g15bqtyidZs
SVM Demo

- [https://cs.stanford.edu/people/karpathy/svmjs/demo/](https://cs.stanford.edu/people/karpathy/svmjs/demo/)
Support Vector Machine

• Cost function

• Large margin classification

• Kernels

• Using an SVM
Using SVM

- SVM software package (e.g., liblinear, libsvm) to solve for $\theta$

- Need to specify:
  - Choice of parameter $C$.
  - Choice of kernel (similarity function):

- Linear kernel: Predict $y = 1$ if $\theta^\top x \geq 0$

- Gaussian kernel:
  - $f_i = \exp(-\frac{\|x-l(i)\|^2}{2\sigma^2})$, where $l(i) = x(i)$
  - Need to choose $\sigma^2$. Need proper feature scaling
Kernel (similarity) functions

• Note: not all similarity functions make valid kernels.

• Many off-the-shelf kernels available:
  • Polynomial kernel
  • String kernel
  • Chi-square kernel
  • Histogram intersection kernel
Multi-class classification

- Use one-vs.-all method. Train $K$ SVMs, one to distinguish $y = i$ from the rest, get $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}$

- Pick class $i$ with the largest $\theta^{(i)}^T x$

Slide credit: Andrew Ng
Logistic regression vs. SVMs

• \( n \) = number of features \( (x \in \mathbb{R}^{n+1}) \), \( m \) = number of training examples

1. **If \( n \) is large (relative to \( m \)):** \((n = 10,000, m = 10 - 1000)\)  
   → Use logistic regression or SVM without a kernel ("linear kernel")

2. **If \( n \) is small, \( m \) is intermediate:** \((n = 1 - 1000, m = 10 - 10,000)\)  
   → Use SVM with Gaussian kernel

3. **If \( n \) is small, \( m \) is large:** \((n = 1 - 1000, m = 50,000+)\)  
   → Create/add more features, then use logistic regression of linear SVM

Neural network likely to work well for most of these case, but slower to train

Slide credit: Andrew Ng
Things to remember

• Cost function

\[
\min_\theta C \left[ \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T f^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T f^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{m} \theta_j^2
\]

• Large margin classification

• Kernels

• Using an SVM