Lecture 15:
- Support Vector Machines
Administrative

• We will have a make-up lecture on Saturday April 23, 2016.

• Project progress reports are due April 21, 2016!
Midterm Exam Statistics

Average: 35.25
Std.dev: 11.31
Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Today

• Support Vector Machines
  - Large Margin Separation
  - Optimization Problem
  - Support Vectors
Recap: Binary Classification Problem

- Training data: sample drawn i.i.d. from set $X \subseteq \mathbb{R}^N$ according to some distribution $D$,

  \[ S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in X \times \{-1, +1\}. \]

- Problem: find hypothesis $h : X \mapsto \{-1, +1\}$ in $H$ (classifier) with small generalization error $R_D(h)$.

- Linear classification:
  - Hypotheses based on hyperplanes.
  - Linear separation in high-dimensional space.
Example: Spam

• Imagine 3 features (spam is “positive” class):
  1. free (number of occurrences of “free”)
  2. money (occurrences of “money”)
  3. BIAS (intercept, always has value 1)

\[
\sum_i w_i \cdot f_i(x) = w \cdot f(x) > 0 \rightarrow \text{SPAM!!!}
\]
Binary Decision Rule

- In the space of feature vectors
  - Examples are points
  - Any weight vector is a hyperplane
  - One side corresponds to $Y = +1$
  - Other corresponds to $Y = -1$

\[ f \cdot w = 0 \]

-1 = HAM
+1 = SPAM

BIAS : -3
free : 4
money : 2
...
The perceptron algorithm

- Start with weight vector $= 0$
- For each training instance $(x_i, y_i^*)$:
  - Classify with current weights
    $$y_i = \begin{cases} 
      +1 & \text{if } w \cdot f(x_i) \geq 0 \\
      -1 & \text{if } w \cdot f(x_i) < 0 
    \end{cases}$$
  - If correct (i.e. $y = y_i^*$), no change!
  - If wrong: update
    $$w = w + y_i^* f(x_i)$$
Linearly separable data

\[ \exists w \text{ such that } \forall t \quad y_t(w \cdot x_t) \geq \gamma > 0 \]

called the margin

Equivalently, for \( y_t = +1 \),
\[ w \cdot x_t \geq \gamma \]

And for \( y_t = -1 \)
\[ w \cdot x_t \leq -\gamma \]
Properties of the perceptron algorithm

• **Separability**: some parameters get the training set perfectly correct

• **Convergence**: if the training is linearly separable, perceptron will eventually converge
Problems with the perceptron algorithm

- **Noise:** if the data isn’t linearly separable, no guarantees of convergence or accuracy

- Frequently the training data is linearly separable! **Why?**
  - When the number of features is much larger than the number of data points, there is lots of flexibility
  - As a result, Perceptron can significantly overfit the data

- **Averaged** perceptron is an algorithmic modification that helps with both issues
  - Averages the weight vectors across all iterations

slide by David Sontag
Linear Separators

- Which of these linear separators is optimal?
Support Vector Machines
Linear Separator

Ham

Spam
Large Margin Classifier
Review: Normal to a plane

\[ w \cdot x + b = 0 \]

\[ \frac{w}{\|w\|} \] -- unit vector normal to \( w \)

\[ \bar{x}_j \] -- projection of \( x_j \)

onto the plane

\[ x_j - \bar{x}_j = \lambda \frac{w}{\|w\|} \]

\( \lambda \) is the length of the vector, i.e.

\[ x_j - \bar{x}_j = \frac{\lambda}{\|w\|} \|w\| = \lambda \]
Any other ways of writing the same dividing line?

- \( w \cdot x + b = 0 \)
- \( 2w \cdot x + 2b = 0 \)
- \( 1000w \cdot x + 1000b = 0 \)
- ....
During learning, we set the scale by asking that, for all $t$,

for $y_t = +1$, $w \cdot x_t + b \geq 1$

and for $y_t = -1$, $w \cdot x_t + b \leq -1$

That is, we want to satisfy all of the **linear** constraints

$$y_t(w \cdot x_t + b) \geq 1 \quad \forall t$$
Large Margin Classifier

\[ \langle w, x \rangle + b \leq -1 \]

\[ \langle w, x \rangle + b \geq 1 \]

linear function

\[ f(x) = \langle w, x \rangle + b \]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \frac{\langle x_+ - x_-, w \rangle}{2 \| w \|} = \frac{1}{2 \| w \|} \left[ [\langle x_+, w \rangle + b] - [\langle x_-, w \rangle + b] \right] = \frac{1}{\| w \|} \]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \langle w, x \rangle + b = 1 \]

Optimization problem

\[
\max_{w,b} \frac{1}{\|w\|} \quad \text{subject to } y_i \left[ \langle x_i, w \rangle + b \right] \geq 1
\]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \langle w, x \rangle + b = 1 \]

Optimization problem

\[ \text{minimize} \quad \frac{1}{2} \| w \|^2 \quad \text{subject to} \quad y_i \left[ \langle x_i, w \rangle + b \right] \geq 1 \]
Convex Programs for Dummies

• Primal optimization problem

\[
\text{minimize } f(x) \text{ subject to } c_i(x) \leq 0
\]

• Lagrange function

\[
L(x, \alpha) = f(x) + \sum_i \alpha_i c_i(x)
\]

• First order optimality conditions in \( x \)

\[
\partial_x L(x, \alpha) = \partial_x f(x) + \sum_i \alpha_i \partial_x c_i(x) = 0
\]

• Solve for \( x \) and plug it back into \( L \)

\[
\text{maximize } L(x(\alpha), \alpha)
\]

(keep explicit constraints)
Dual Problem

• Primal optimization problem

\[
\text{minimize } \frac{1}{2} \|w\|^2 \text{ subject to } y_i [\langle x_i, w \rangle + b] \geq 1
\]

• Lagrange function

\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i [y_i [\langle x_i, w \rangle + b] - 1]
\]

Optimality in \( w, b \) is at saddle point with \( \alpha \)

• Derivatives in \( w, b \) need to vanish
Dual Problem

- Lagrange function
  \[ L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i [y_i \langle x_i, w \rangle + b] - 1 \]

- Derivatives in \( w, b \) need to vanish
  \[ \partial_w L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0 \]
  \[ \partial_b L(w, b, a) = \sum_i \alpha_i y_i = 0 \]

- Plugging terms back into \( L \) yields
  \[
  \text{maximize } - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \\
  \text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0
  \]
Support Vector Machines

\[
\begin{align*}
\minimize_{w,b} \quad & \frac{1}{2} \|w\|^2 \\
\text{subject to} \quad & y_i \left[ \langle x_i, w \rangle + b \right] \geq 1
\end{align*}
\]

\[
w = \sum_{i} y_i \alpha_i x_i
\]

\[
\begin{align*}
\maximize_{\alpha} \quad & -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i} \alpha_i \\
\text{subject to} \quad & \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0
\end{align*}
\]
minimize $\frac{1}{2} \| w \|^2$ subject to $y_i [\langle x_i, w \rangle + b] \geq 1$

\[ w = \sum_i y_i \alpha_i x_i \]

Karush Kuhn Tucker Optimality condition
\[ \alpha_i [y_i [\langle w, x_i \rangle + b] - 1] = 0 \]
\[ \alpha_i = 0 \]
\[ \alpha_i > 0 \iff y_i [\langle w, x_i \rangle + b] = 1 \]
• Weight vector $w$ as weighted linear combination of instances
• Only points on margin matter (ignore the rest and get same solution)
• Only inner products matter
  – Quadratic program
  – We can replace the inner product by a kernel
• Keeps instances away from the margin
Example
Example

Number of Support Vectors: 3  (-ve: 2, +ve: 1)  Total number of points: 15
Why Large Margins?

- Maximum robustness relative to uncertainty
- Symmetry breaking
- Independent of correctly classified instances
- Easy to find for easy problems