Linear classifiers – Which line is better?
Pick the one with the largest margin!

confidence = \( y^j (w \cdot x^j + w_0) \)

Maximize the margin

maximize worst case margin:

\[
\begin{align*}
\max_{\gamma,w,w_0} \min_{y^j} y^j (w \cdot x^j + w_0) \\
\text{s.t.} & \quad y^j (w \cdot x^j + w_0) \geq \gamma, \forall j \in \{1, \ldots, N\}
\end{align*}
\]
But there are many planes...

$w \cdot x + w_0 = 0$

Review: Normal to a plane

$x_j^* = \bar{x}_j + \lambda \frac{w}{||w||}$

$\lambda$ will increase in magnitude as we scale $w/||w||$.

Insensitive to scaling.

Projection onto plane.
A Convention: Normalized margin –
Canonical hyperplanes

$$x_i^- = x_i^+ + \frac{\gamma}{\|w\|}$$

Margin maximization using canonical hyperplanes

Unnormalized problem:

$$\max_{\gamma, w, w_0} \gamma \quad \text{s.t.} \quad y_j (w \cdot x_j^+ + w_0) \geq \gamma, \forall j \in \{1, \ldots, N\}$$

Normalized Problem:

$$\min_{w, w_0} \|w\|^2_2 \quad \text{s.t.} \quad y_j (w \cdot x_j^+ + w_0) \geq 1, \forall j \in \{1, \ldots, N\}$$
Support vector machines (SVMs)

\[
\min_{\mathbf{w}, \mathbf{w}_0} \frac{1}{2} \| \mathbf{w} \|^2 \\
y^j (\mathbf{w} \cdot \mathbf{x}_j + w_0) \geq 1, \forall j \in \{1, \ldots, N\}
\]

- Solve efficiently by many methods, e.g.,
  - quadratic programming (QP)
  - Well-studied solution algorithms
  - Stochastic gradient descent

Hyperplane defined by support vectors

Only perturbations of few points will change solution

What if the data is not linearly separable?

Use features of features of features of features...

\[
\phi(x) = \begin{pmatrix} x^2 \\ x \\ e^x \sin x \\ \vdots \end{pmatrix}
\]

Can be done efficiently with kernels
(Some popularized kernels in ML)
What if the data is still not linearly separable?

If data is not linearly separable, some points don’t satisfy margin constraint:

\[ y_j (w \cdot x_j + w_0) \geq 1, \forall j \]

How bad is the violation?

\[ \text{Margin violation} = \begin{cases} 0, & 1 - y_j (w \cdot x_j + w_0) > 0 \\ 1 - y_j (w \cdot x_j + w_0), & \text{otherwise} \end{cases} \]

Tradeoff margin violation with \( ||w|| \):

\[
\min_{w, w_0} ||w||^2 \text{ of this constraint}
\]

\[
\sum_{j=1}^{N} (-y_j (w \cdot x_j + w_0))^+ + \text{regularization}
\]

SVMs for Non-Linearily Separable meet my friend the Perceptron...

Perceptron was minimizing the hinge loss:

\[
\sum_{j=1}^{N} (-y_j (w \cdot x_j + w_0))^+ + \text{regularization}
\]

SVMs minimizes the regularized hinge loss!!

\[
||w||^2 + C \sum_{j=1}^{N} (1 - y_j (w \cdot x_j + w_0))^+
\]

both hinge loss

just convention
Stochastic Gradient Descent for SVMs

- Perceptron minimization:
  \[ \sum_{j=1}^{N} (-y_j(w \cdot x_j + w_0))_+ \]
- SGD for Perceptron:
  \[ w(t+1) \leftarrow w(t) + \eta \left( y(t)(w(t) \cdot x(t)) \right) \]
  \[ \text{step size } \eta \]

- SVMs minimization:
  \[ \|w\|^2 + C \sum_{j=1}^{N} (1 - y_j(w \cdot x_j + w_0))_+ \]
- SGD for SVMs:
  \[ w(t+1) \leftarrow w(t) + \eta \left( C \left( 1 - y(t)(w(t) \cdot x(t)) \right) y(t)x(t) \right) \]
  \[ \text{step size } \eta \]

What you need to know

- Maximizing margin
- Derivation of SVM formulation
- Non-linearly separable case
  - Hinge loss
  - A.K.A. adding slack variables
- SVMs = Perceptron + L2 regularization
- Can optimize SVMs with SGD
  - Many other approaches possible
What you have learned thus far

- Learning is function approximation
- Point estimation
- Regression
- LASSO
- Logistic regression
- Bias-Variance tradeoff
- Regularization
- Decision trees
- Cross validation
- Boosting
- Instance-based learning
- Online learning
- Perceptron
- SVMs
- Kernel trick
Review material in terms of...

- Types of learning problems
- Hypothesis spaces
- Loss functions
- Optimization algorithms

ML Pipeline

<table>
<thead>
<tr>
<th>Attributes/Observations</th>
<th>Features/Basis Functions</th>
<th>Task</th>
<th>Hypothesis Class/Model</th>
<th>Algorithm/Optimization Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>h_i(x)</td>
<td></td>
<td>Linear models w, x</td>
<td>optimize a loss function</td>
</tr>
<tr>
<td>age</td>
<td>φ(x)</td>
<td></td>
<td>NN, DTs</td>
<td>gradient</td>
</tr>
<tr>
<td>gender</td>
<td>(age, p, w, b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td></td>
<td></td>
<td></td>
<td>set gradient = 0</td>
</tr>
<tr>
<td>p, w, b</td>
<td></td>
<td></td>
<td></td>
<td>closed form</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Stochastic Gradient descent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SGLD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>search decision tree</td>
</tr>
</tbody>
</table>
Learning Task/Measuring Error

<table>
<thead>
<tr>
<th>TASK</th>
<th>LOSS FUNCTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Squared error</td>
</tr>
</tbody>
</table>
| Classification      | log loss for logistic regression, 
                       | hinge loss     |
| Density Estimation  | log loss for LR  |

Hypothesis Classes & Decision Boundaries

Simple Linear Model

Linear Model with Higher-Order Features or Kernels

Nearest Neighbors

Boosting
The Power of Regularization

Overfitting
Bias/Variance Tradeoff

Regularization

\[ \|W\|_2^2 \leq L_2 \text{ regularizer} \]

Your Midterm…

- Content: Everything up to today…
- Only 50mins, so arrive early and settle down quickly
- “Open book”
  - Textbook, Course notes, Personal notes
- No:
  - Computer, phone, other materials,…
- The exam:
  - Covers key concepts and ideas, work on understanding the big picture, and differences between methods