CS142: Machine Learning
Lecture 3: Classification & ROC Curves, Directed Graphical Models

Instructor: Erik Sudderth
Brown University Computer Science
September 22, 2015

Figure credits:
PRML: Pattern Recognition & Machine Learning, Bishop 2007
ESL: Elements of Statistical Learning, Hastie & Tibshirani & Friedman 2009
CS142 Announcements

- **Homework 1**
  Due Thursday 9/24 at 11:59pm.
  Be sure to read the collaboration & grading policy document!

- **Electronic Submission of Homeworks**
  Detailed instructions were announced and posted online.
  cs142student group was setup today, handin should now work.
  To submit from outside the CIT you must setup ssh in advance!
  You can run the handin script multiple times. Test before the deadline!
CS142 ML: Lecture 2 Outline

- Decision Theory & ROC Curves
- Directed graphical models
Binary Classification Problems

\[ x \in \mathbb{R}^d \]  
\textit{d-dimensional feature vector for each example}

\[ t \in \{0, 1\} \]  
\textit{discrete class/category label}

\[ p(x, t) \]  
\textit{true joint probability distribution (unknown)}

Training Data:

\[ (x_n, t_n), \quad n = 1, 2, \ldots, N \]

\( N \) independent & identically distributed (iid) samples from \( p(x, t) \)

Classification Goal:

Learn a function \( y(x) \in \{0, 1\} \) for which \( P(y(x) \neq t) \approx 0 \)
Bayesian Decision Theory

We are given both a **probabilistic model** and a **loss function**:

**Prior distribution:**

\[ P(t = 0) = q \quad P(t = 1) = 1 - q \]

**Likelihood function:**

\[ p(x \mid t) \]

**Posterior distribution:**

\[ p(t \mid x) = \frac{p(x \mid t)p(t)}{p(x)} \]

**Loss function:**

\[ L(t, y) = \text{cost of predicting class } y \text{ when } t \text{ is true.} \]

\[ L(0, 1) = \lambda_{01} > 0 \quad L(1, 0) = \lambda_{10} > 0 \]

\[ L(0, 0) = L(1, 1) = 0 \]

Coming soon: many examples!
Likelihood Ratio Classifiers

The optimal decision minimizes the posterior expected loss:

$$y(x) = \arg\min_y E[L(t, y) \mid X = x] = \arg\min_y \sum_{t=0}^1 L(t, y)p(t \mid x)$$

It is optimal to decide $t=1$ if and only if:

$$\frac{p(x \mid t = 1)}{p(x \mid t = 0)} \geq \left( \frac{q}{1 - q} \right) \cdot \left( \frac{\lambda_{01}}{\lambda_{10}} \right)$$

**maximum a posteriori (MAP) rule**

$$\frac{p(x \mid t = 1)}{p(x \mid t = 0)} \geq \left( \frac{q}{1 - q} \right) \quad \text{if } \lambda_{01} = \lambda_{10}$$

**maximum likelihood (ML) rule**

$$\frac{p(x \mid t = 1)}{p(x \mid t = 0)} \geq 1 \quad \text{if } q = 0.5 \text{ and } \lambda_{01} = \lambda_{10}$$
### False Positives versus False Negatives

- **False positive (FP):** Predict class 1 when truth is class 0
- **False negative (FN):** Predict class 0 when truth is class 1
- **True positive (TP):** Predict class 1 when truth is class 1
- **True negative (TN):** Predict class 0 when truth is class 0

<table>
<thead>
<tr>
<th>Truth</th>
<th>1</th>
<th>0</th>
<th>( \Sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>TP</td>
<td>FP</td>
<td>( \hat{N}_+ = TP + FP )</td>
</tr>
<tr>
<td>0</td>
<td>FN</td>
<td>TN</td>
<td>( \hat{N}_- = FN + TN )</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>( N_+ = TP + FN )</td>
<td>( N_- = FP + TN )</td>
<td>( N = TP + FP + FN + TN )</td>
</tr>
</tbody>
</table>

- **Sensitivity, recall, or true positive rate (TPR)**
- **False alarm rate or false positive rate (FPR)**

\[
\text{TPR} = \frac{TP}{N_+} \approx P(y(x) = 1 \mid t = 1) \quad \text{FPR} = \frac{FP}{N_-} \approx P(y(x) = 1 \mid t = 0)
\]

- **Receiver operating characteristic (ROC):** Plot of TPR vs FPR
Idealized ROC Curves

Set of log-likelihood ratio classifiers:

\[
\log \frac{p(x \mid t = 1)}{p(x \mid t = 0)} > \tau
\]

Interesting points:
- Perfect classifier (may be impossible)
- Always select 1: \( \tau \rightarrow -\infty \)
- Always select 0: \( \tau \rightarrow +\infty \)
- Randomized classifiers (ignore data)
- What if TPR < FPR? Can we do better?
Homework explores in detail!

\[
TPR = \frac{TP}{N_+} \approx P(y(x) = 1 \mid t = 1)
\]
\[
FPR = \frac{FP}{N_-} \approx P(y(x) = 1 \mid t = 0)
\]

**EER:** Equal Error Rate

**AUC:** Area Under Curve
Choosing a point on ROC curve

Set of log-likelihood ratio classifiers:

\[
\log \frac{p(x \mid t = 1)}{p(x \mid t = 0)} > \tau
\]

Bayesian decision theoretic optimum:

\[
\tau = \log \left( \frac{q}{1 - q} \cdot \frac{\lambda_{01}}{\lambda_{10}} \right)
\]

- If our data-generation model is perfectly correct, provably better than any other possible classifier!
- In practice we learn approximate models from data, compare via ROC
Choosing a point on ROC curve

Set of log-likelihood ratio classifiers:

\[
\log \left( \frac{p(x \mid t = 1)}{p(x \mid t = 0)} \right) > \tau
\]

Neyman-Pearson Classifier Design:

- Upper bound false alarm rate by \( \alpha \)
- Seek to maximize true positive rate (power):
  \[
  \max \ TPR \text{ subject to } FPR \leq \alpha
  \]
- Optimal classifier is always a likelihood ratio test, for a threshold \( \tau \) derived from \( \alpha \)
  (Proof: See CS145 lectures & readings)

\[TPR = \frac{TP}{N_+} \approx P(y(x) = 1 \mid t = 1)\]

\[FPR = \frac{FP}{N_-} \approx P(y(x) = 1 \mid t = 0)\]
The number of *negative* examples may not be well defined:
- How many windows not containing a car are there in an image?
- How many documents not about cars exist in the world?
Idealized Precision-Recall Curves

Recall:
\[
\frac{TP}{N_+} \approx p(y(x) = 1 \mid t = 1)
\]

Precision:
\[
\frac{TP}{\hat{N}_+} \approx p(t = 1 \mid y(x) = 1)
\]
Precision-Recall for Object Detection

Recall:
\[
\frac{TP}{N_+} \approx p(y(x) = 1 \mid t = 1)
\]

Precision:
\[
\frac{TP}{\hat{N}_+} \approx p(t = 1 \mid y(x) = 1)
\]

Fei-Fei, Fergus, & Torralba, ICCV 2009
Decision Theory & ROC Curves
Directed graphical models
Directed Graphical Models

Chain rule implies that any joint distribution equals:
\[ p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1) \cdots p(x_N \mid x_{N-1}, \ldots, x_1) \]

Directed graphical model implies a restricted factorization:
\[ p(x) = \prod_{s=1}^{N} p(x_s \mid x_{\Gamma(s)}) \]
\[ \Gamma(s) \rightarrow \text{set of parents of node } s, \text{possibly empty} \]

Valid for any directed acyclic graph (DAG):
equivalent to dropping conditional dependencies in chain rule

\[ p(x_{1:5}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)p(x_5 \mid x_1, x_2, x_3, x_4) \]
\[ = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2, x_3)p(x_5 \mid x_3) \]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Choosing an ordering
Without loss of generality, we can write
\[ p(A, R, E, B) = p(A|R, E, B)p(R, E, B) \]
\[ = p(A|R, E, B)p(R|E, B)p(E, B) \]
\[ = p(A|R, E, B)p(R|E, B)p(E|B)p(B) \]

Assumptions:
- The alarm is not directly influenced by any report on the radio,
  \[ p(A|R, E, B) = p(A|E, B) \]
- The radio broadcast is not directly influenced by the burglar variable,
  \[ p(R|E, B) = p(R|E) \]
- Burglaries don’t directly ‘cause’ earthquakes, \[ p(E|B) = p(E) \]

Therefore
\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \quad p(E = 1) = 0.000001 \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radio = 1</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Why don’t we all use expert systems? Accurate assessment of such probabilities can be hard.
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \quad p(E = 1) = 0.000001 \]

\[ p(A|B, E) \quad p(R|E) \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radio = 1</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Inference: Probability of Burglary given Alarm

\[ p(B = 1|A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \]

\[ = \frac{\sum_{E,R} p(A = 1|B = 1, E)p(B = 1)p(E)p(R|E)}{\sum_{B,E,R} p(A = 1|B, E)p(B)p(E)p(R|E)} \approx 0.99 \]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \quad \quad p(E = 1) = 0.000001 \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Radio = 1 | Earthquake
---------|----------
1         | 1        
0         | 0        

Inference: Probability of Burglary given Alarm and Radio

\[ p(B = 1|A = 1, R = 1) \approx 0.01 \]

“Explaining Away”
Real Alarm Networks: ICU Monitoring

Beinlich et al., 1989

Aleks, Russell, et al., 2008
Reminder: Independent Variables

\[ p(x, y) = \]

\[ X \perp Y \]

\[ p_{XY}(x, y) = p_X(x)p_Y(y) \]

for all \( x \in \mathcal{X}, y \in \mathcal{Y} \)

- Equivalent conditions on conditional probabilities:
  \[ p_{X|Y}(x \mid y) = p_X(x) \] for all \( p_Y(y) > 0 \)
  \[ p_{Y|X}(y \mid x) = p_Y(y) \] for all \( p_X(x) > 0 \)
Reminder: Conditional Independence

\[ P(x, y) = p_{X|Z}(x | z)p_{Y|Z}(y | z) \]

for all \( x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z} \)
Conditional Independence in Directed Models

All belief networks with three nodes and two links:

\begin{align*}
\text{(a)} & \quad p(A, B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C) \\
\text{(b)} & \quad p(A, B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A,C)p(B|C)}{p(C)} = p(A|C)p(B|C) \\
\text{(c)} & \quad p(A, B|C) = \frac{p(A|C)p(C|B)p(B)}{p(C)} = \frac{p(A|C)p(B,C)}{p(C)} = p(A|C)p(B|C)
\end{align*}

In (d) the variables $A, B$ are conditionally dependent given $C$, $p(A, B|C) \propto p(C|A, B)p(A)p(B)$.
In (a), (b) and (c), the variables $A, B$ are marginally dependent.

In (d) the variables $A, B$ are marginally independent.

\[ p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B) \]
Microsoft Trueskill Ranking System

This is a slightly different type of graphical model, called a “factor graph”. May also represent as a directed graphical model.

Herbrich, Minka, & Graepel, NIPS 2006 (and later)