Overview

- Cost-sensitive classification
- Conditional probability estimation
- Fairness

Scoring functions in general

- Statistical model: $(X,Y) \sim P$ for distribution $P$ over $X \times \{-1, +1\}$
- Binary classifiers are generally of the form
  \[ x \mapsto \text{sign}(h(x)) \]
  for some scoring function $h : X \to \mathbb{R}$
  - E.g. Bayes classifier uses scoring function $h(x) = \eta(x) - 1/2$ where $\eta(x) = \Pr(Y = +1 \mid X = x)$
- Use with (surrogate) loss functions like $\ell_{zo}$, $\ell_{log}$, $\ell_{msq}$, $\ell_{hinge}$
  \[ R(h) = \mathbb{E}[\ell(Y h(X))] \]
- Issues to consider:
  - Different types of mistakes have different costs
  - How to get $\Pr(Y = +1 \mid X = x)$ from $h(x)$?
  - More than two classes

Cost-sensitive classification

- **Cost matrix** for different kinds of mistakes (for $c \in [0, 1]$)

<table>
<thead>
<tr>
<th></th>
<th>$\hat{y} = -1$</th>
<th>$\hat{y} = +1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = -1$</td>
<td>0</td>
<td>$c$</td>
</tr>
<tr>
<td>$y = +1$</td>
<td>$1 - c$</td>
<td>0</td>
</tr>
</tbody>
</table>

(Why can we restrict attention to $c \in [0, 1]$?)

- **Cost-sensitive $\ell$-loss**:
  \[ \ell^{(c)}(y, \hat{y}) = \left( \mathbb{1}_{\{y=+1\}} \cdot (1 - c) + \mathbb{1}_{\{y=-1\}} \cdot c \right) \cdot \ell(y \hat{y}) \]

- If $\ell$ is convex in $\hat{y}$, then so is $\ell^{(c)}(y, \cdot)$
- **Cost-sensitive (empirical) risk**:
  \[ R^{(c)}(h) := \mathbb{E}[\ell^{(c)}(Y, h(X))] \]
  \[ \widehat{R}^{(c)}(h) := \frac{1}{n} \sum_{i=1}^{n} \ell^{(c)}(y_i, h(x_i)) \]
Minimizing cost-sensitive risk

- Analogue of Bayes classifier for cost-sensitive (zero-one loss) risk?
- Let \( \eta(x) = \Pr(Y = 1 | X = x) \)
- Fix \( x \); what is conditional cost-sensitive risk of predicting \( \hat{y} \)?

\[
(1 - c) \cdot \eta(x) \cdot \mathbb{1}_{\{\hat{y} = -1\}} + c \cdot (1 - \eta(x)) \cdot \mathbb{1}_{\{\hat{y} = +1\}}.
\]

- Minimized when

\[
\hat{y} = \begin{cases} 
+1 & \text{if } \eta(x) \cdot (1 - c) > (1 - \eta(x)) \cdot c \\
-1 & \text{otherwise}
\end{cases}
\]

- Same as

\[
\hat{y} = \begin{cases} 
+1 & \text{if } \eta(x) > c \\
-1 & \text{otherwise}
\end{cases}
\]

- I.e., use \( \eta \) as scoring function, but threshold at \( c \) instead of 1/2

- Where does \( c \) come from?

Example: balanced error rate

- \textbf{Balanced error rate}: \( \text{BER} := \frac{1}{2} \text{FNR} + \frac{1}{2} \text{FPR} \)
- Which cost sensitive risk to try to minimize?

\[
2 \cdot \text{BER} = \Pr(h(X) \leq 0 | Y = +1) + \Pr(h(X) > 0 | Y = -1) = \frac{1}{\pi} \cdot \Pr(h(X) \leq 0 \wedge Y = +1) + \frac{1}{1 - \pi} \cdot \Pr(h(X) > 0 \wedge Y = -1)
\]

where \( \pi = \Pr(Y = +1) \).

- Therefore, we want to use the following cost matrix:

\[
\begin{array}{c|cc}
\hat{y} & -1 & +1 \\
\hline
y &=& 0 \\
-1 & \frac{1}{\pi} & 0
\end{array}
\]

- This corresponds to \( c = \pi \).

Importance-weighted risk

- Perhaps the world tells you how important each example is
- Statistical model: \((X,Y,W) \sim P\)
  - \( W \) is (non-negative) \textit{importance weight} of example \((X,Y)\)
- \textbf{Importance-weighted }\ell\textbf{-risk} of \( h \):

\[
\mathbb{E}[W \cdot \ell(Yh(X))]
\]

- Estimate from data \((x_1, y_1, w_1), \ldots, (x_n, y_n, w_n)\):

\[
\frac{1}{n} \sum_{i=1}^{n} w_i \cdot \ell(y_i h(x_i))
\]

- Comes up in many machine learning applications (e.g., boosting)

Conditional probability estimation I

- How to get estimate of \( \eta(x) = \Pr(Y = +1 | X = x) \)?
- Useful if want to know expected cost of a prediction

\[
\mathbb{E}[\ell^{(c)}(Yh(X)) | X = x] = \begin{cases} 
(1 - c) \cdot \eta(x) & \text{if } h(x) \leq 0 \\
c \cdot (1 - \eta(x)) & \text{if } h(x) > 0
\end{cases}
\]

- Squared loss risk minimized by scoring function

\[
h(x) = 2\eta(x) - 1.
\]

Therefore, given \( h \), can estimate \( \eta \) using \( \hat{\eta}(x) = \frac{1 + h(x)}{2} \)

- Recipe:
  - Find scoring function \( h \) that (approximately) minimizes (empirical) squared loss risk
  - Construct conditional probability estimate \( \hat{\eta} \) using above formula
Conditional probability estimation II

▶ Similar strategy available for logistic loss
▶ But not for hinge loss!
  ▶ Hinge loss risk is minimized by \( h(x) = \text{sign}(2\eta(x) - 1) \)
  ▶ Cannot recover \( \eta \) from \( h \)
▶ Caveat: If using insufficiently expressive functions for \( h \) (e.g., linear functions), may be far from minimizing squared loss risk
  ▶ Fix: use more flexible models (e.g., feature expansion, neural nets)

Application: Reducing multi-class to binary

▶ Multi-class: Conditional probability function is vector-valued function

\[
\eta(x) = \begin{bmatrix}
\Pr(Y = 1 \mid X = x) \\
\vdots \\
\Pr(Y = K \mid X = x)
\end{bmatrix}
\]

▶ Reduction: learn \( K \) scalar-valued functions, the \( k \)-th function is supposed to approximate \( \eta_k(x) = \Pr(Y = k \mid X = x) \).
  ▶ This can be done by create \( K \) binary classification problems, where in problem \( k \), label is \( I_{\{y=k\}} \).
  ▶ Given the \( K \) learned conditional probability functions \( \hat{\eta}_1, \ldots, \hat{\eta}_K \), we form a final predictor \( \hat{f} \)

\[
\hat{f}(x) = \arg \max_{k=1,\ldots,K} \hat{\eta}_k(x).
\]

When does one-against-all work well?

▶ If learned conditional probability functions \( \hat{\eta}_k \) are accurate, then behavior of one-against-all classifier \( \hat{f} \) is similar to optimal classifier

\[
f^*(x) = \arg \max_{k=1,\ldots,K} \Pr(Y = k \mid X = x).
\]

▶ Claim:

\[
\text{err}(\hat{f}) \leq \text{err}(f^*) + 2 \cdot \mathbb{E} \left[ \max_k |\hat{\eta}_k(X) - \eta_k(X)| \right].
\]

Fairness

▶ Use of predictive models (e.g., in hiring, criminal justice) has raised concerns about whether they offer "fair treatment" to individuals and/or groups
  ▶ We will focus on group-based fairness
  ▶ Individual-based fairness also important, but not as well-studied
**Disparate treatment**

- Often predictive models work better for some groups than for others
  - Example: face recognition (Buolamwini and Gebru, 2018; Lohr, 2018)

**Some possible causes**

- People deliberately being unfair
- Disparity in number of available training data for different groups
- Disparity in usefulness of available features for different groups
- Disparity in relevance of prediction problem for different groups
  - ...

**ProPublica study**

- ProPublica studied predictive model being used to determine “pre-trial detention” (Angwin et al, 2016)
  - Judge needs to decide whether or not an arrested defendant should be released while awaiting trial
  - Predictive model (“COMPAS”) provides an estimate of \( \Pr(Y = 1 \mid X = x) \) where \( Y = \mathbb{1} \) (will commit (violent) crime if released) and \( X \) is “features” of defendant.
  - Study argued that COMPAS treated black defendants unfairly in a certain sense

**Group fairness criteria**

- Setup:
  - \( X \): features for individual
  - \( A \): group membership attribute (e.g., race, sex, age, religion)
  - \( Y \): outcome variable (e.g., “will repay loan”, “will re-offend”)
  - \( \hat{Y} \): prediction of outcome variable (as function of \((X, A))\)
  - For simplicity, assume \( A, Y, \hat{Y} \) are \(\{0, 1\}\)-valued
- Many fairness criteria are based on joint distribution of \((A, Y, \hat{Y})\)
- Caveat: Often, we don’t have access to \(Y\) in training data
Classification parity

- Fairness criterion: *Classification parity*

\[ \Pr(\hat{Y} = 1 \mid A = 0) \approx \Pr(\hat{Y} = 1 \mid A = 0) \]

- Sounds reasonable, but easy to satisfy with perverse methods
- Example: trying to predict \( Y = 1 \{\text{will repay loan if given one}\} \)
- Suppose conditional distributions of \((Y, \hat{Y})\) given \( A \) are as follows:

<table>
<thead>
<tr>
<th>( A = 0 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
<th>( A = 1 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.27</td>
<td>0.22</td>
<td>0.46</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.14</td>
<td>0.37</td>
<td>0.19</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

- For \( A = 0 \) people, correctly give out loans to people who will repay
- For \( A = 1 \) people, give out loans randomly
- Satisfies criterion, but bad for \( A = 1 \) people

Equalized odds I

- Fairness criterion: *Equalized odds*

\[ \Pr(\hat{Y} = 1 \mid Y = y, A = 0) \approx \Pr(\hat{Y} = 1 \mid Y = y, A = 1) \]

for both \( y \in \{0, 1\} \).
- In particular, FPR and FNR must be about the same across groups.
- Could also just ask for *Equalized FPR*, or *Equalized FNR*
- Previous example violates Equalized FPR:

<table>
<thead>
<tr>
<th>( A = 0 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
<th>( A = 1 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.27</td>
<td>0.22</td>
<td>0.46</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.14</td>
<td>0.37</td>
<td>0.19</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

E.g., \( A = 0 \) group has 0% FPR, while \( A = 1 \) group has 50% FPR.

Equalized odds II

- ProPublica study:
  - Found that FPR for \( A = 0 \) group (black defendants; 45%) was higher than FPR for \( A = 0 \) group (white defendants; 23%)

<table>
<thead>
<tr>
<th>( A = 0 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
<th>( A = 1 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.27</td>
<td>0.22</td>
<td>0.46</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.14</td>
<td>0.37</td>
<td>0.19</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

\[ \frac{0.22}{0.27 + 0.22} \approx 0.45, \quad \frac{0.14}{0.46 + 0.14} \approx 0.23 \]

- Arrest a lot of college protesters from \( A = 0 \) group
  - Most college protesters are unlikely to (violently) (re-)offend (COMPAS correctly predicts as such), so decreases denominator in FPR estimate
  - FPR’s are artificially equalized
- Very hard to get the “right” data sets that can be compared

Critique of group fairness criteria

- Example from (Corbett-Davies and Goel, 2018)
- Can easily manipulate data sets to achieve Equalized FPR

<table>
<thead>
<tr>
<th>( A = 0 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
<th>( A = 1 )</th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.27</td>
<td>0.22</td>
<td>0.46</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.14</td>
<td>0.37</td>
<td>0.19</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

\[ \frac{0.22}{0.27 + 0.22} \approx 0.45, \quad \frac{0.14}{0.46 + 0.14} \approx 0.23 \]

- Arrest a lot of college protesters from \( A = 0 \) group
  - Most college protesters are unlikely to (violently) (re-)offend (COMPAS correctly predicts as such), so decreases denominator in FPR estimate
  - FPR’s are artificially equalized
- Very hard to get the “right” data sets that can be compared