Classification objectives

COMS 4721 Spring 2018
1. Recap: binary classification
Consider binary classification problems with $\mathcal{Y} = \{-1, +1\}$. 
Scoring functions

Consider binary classification problems with $\mathcal{Y} = \{-1, +1\}$.

WLOG, can write a classifier as $x \mapsto \begin{cases} +1 & \text{if } h(x) > 0 \\ -1 & \text{if } h(x) \leq 0 \end{cases}$

for some \textit{scoring function} $h : \mathcal{X} \rightarrow \mathbb{R}$.
Scoring functions

Consider binary classification problems with $\mathcal{Y} = \{-1, +1\}$.

WLOG, can write a classifier as $x \mapsto \begin{cases} +1 & \text{if } h(x) > 0 \\ -1 & \text{if } h(x) \leq 0 \end{cases}$

for some scoring function $h : \mathcal{X} \rightarrow \mathbb{R}$.

Some examples:

For linear classifiers: $h(x) = x^T w$

For Bayes optimal classifier: $h(x) = \mathbb{P}(Y = 1 \mid X = x) - 1/2$
Scoring functions

Consider binary classification problems with $\mathcal{Y} = \{-1, +1\}$.

WLOG, can write a classifier as $x \mapsto \begin{cases} +1 & \text{if } h(x) > 0 \\ -1 & \text{if } h(x) \leq 0 \end{cases}$

for some scoring function $h : \mathcal{X} \to \mathbb{R}$.

Some examples:
For linear classifiers: $h(x) = x^T w$
For Bayes optimal classifier: $h(x) = \mathbb{P}(Y = 1 \mid X = x) - 1/2$

Expected zero-one loss is $\mathbb{E} \left[ \ell_{0/1} \left( Y h(X) \right) \right]$ where $\ell_{0/1}(z) = 1 \{ z \leq 0 \}$:
Zero-one loss and some surrogate losses

Some *surrogate losses* that upper-bound $\ell_{0/1}$:

<table>
<thead>
<tr>
<th>Loss Type</th>
<th>Loss Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>hinge</td>
<td>$\ell_{\text{hinge}}(z) = [1 - z]_+$</td>
</tr>
<tr>
<td>squared</td>
<td>$\ell_{\text{sq}}(z) = (1 - z)^2$</td>
</tr>
<tr>
<td>logistic</td>
<td>$\ell_{\text{logistic}}(z) = \log_2(1 + e^{-z})$</td>
</tr>
<tr>
<td>exponential</td>
<td>$\ell_{\text{exp}}(z) = e^{-z}$</td>
</tr>
</tbody>
</table>
Some surrogate losses that upper-bound \( \ell_{0/1} \):

<table>
<thead>
<tr>
<th>Loss</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>hinge</td>
<td>( \ell_{\text{hinge}}(z) = [1 - z]_+ )</td>
</tr>
<tr>
<td>squared</td>
<td>( \ell_{\text{sq}}(z) = (1 - z)^2 )</td>
</tr>
<tr>
<td>logistic</td>
<td>( \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) )</td>
</tr>
<tr>
<td>exponential</td>
<td>( \ell_{\exp}(z) = e^{-z} )</td>
</tr>
</tbody>
</table>

**Note:** when \( y \in \{\pm 1\} \) and \( p \in \mathbb{R} \),

\[
\ell_{\text{sq}}(yp) = (1 - yp)^2 = (y - p)^2.
\]
Beyond binary classification

- What if different types of mistakes have different costs?
- What if we want the (conditional) probability of a positive label?
- What if there are more than two classes?
2. Cost-sensitive classification
Cost-sensitive classification

Often have **different costs 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 $c \in [0, 1]$?)

**Fact**: if $\ell$ is convex, then so is $\ell(c)(y, \cdot)$ for each $y \in \{\pm 1\}$.

**Cost-sensitive (empirical) $\ell$-risk of scoring function $h: X \to \mathbb{R}$**:

$$R(c)(h) := \mathbb{E}[\ell(c)(Y, h(X))]$$

Our actual objective.

$$\hat{R}(c)(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(c)(Y_i, h(X_i))$$

What we can try to minimize.
Cost-sensitive classification

Often have **different costs 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 classification

Often have **different costs 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 zero-one loss**:

\[
\ell_{0/1}^{(c)}(y, p) = \left( \mathbb{1}\{y = 1\} \cdot (1 - c) + \mathbb{1}\{y = -1\} \cdot c \right) \cdot \ell_{0/1}(yp).
\]
Cost-sensitive classification

Often have *different costs 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* (for loss \( \ell : \mathbb{R} \to \mathbb{R} \)):

\[
\ell^{(c)}(y, p) = (\mathbbm{1}\{y = 1\} \cdot (1 - c) + \mathbbm{1}\{y = -1\} \cdot c) \cdot \ell(yp).
\]
Cost-sensitive classification

Often have **different costs 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** (for loss \( \ell : \mathbb{R} \to \mathbb{R} \)):

\[
\ell^{(c)}(y, p) = (1 \{ y = 1 \} \cdot (1 - c) + 1 \{ y = -1 \} \cdot c) \cdot \ell(yp).
\]

**Fact**: if \( \ell \) is convex, then so is \( \ell^{(c)}(y, \cdot) \) for each \( y \in \{ \pm 1 \} \).
Cost-sensitive classification

Often have **different costs 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** (for loss \( \ell : \mathbb{R} \to \mathbb{R} \)):

\[
\ell^{(c)}(y, \hat{y}) = (1 \{ y = 1 \} \cdot (1 - c) + 1 \{ y = -1 \} \cdot c) \cdot \ell(yp).
\]

**Fact**: if \( \ell \) is convex, then so is \( \ell^{(c)}(y, \cdot) \) for each \( y \in \{ \pm 1 \} \).

**Cost-sensitive (empirical) \( \ell \)-risk** of scoring function \( h : \mathcal{X} \to \mathbb{R} \):

\[
\mathcal{R}^{(c)}(h) := \mathbb{E} \left[ \ell^{(c)}(Y, h(X)) \right]
\]

Our actual objective.

\[
\widehat{\mathcal{R}}^{(c)}(h) := \frac{1}{n} \sum_{i=1}^{n} \ell^{(c)}(Y_i, h(X_i))
\]

What we can try to minimize.
Minimizer of cost-sensitive risk

What is the optimal classifier for cost-sensitive (zero-one loss) risk?

Let $\eta(x) := P(Y = 1 | X = x)$ for $x \in X$.

Therefore, the classifier based on scoring function $h(x) := \eta(x) - c$, $x \in X$ has the smallest cost-sensitive risk $R(c)$.
What is the optimal classifier for cost-sensitive (zero-one loss) risk?

Let $\eta(x) := \mathbb{P}(Y = 1 \mid X = x)$ for $x \in \mathcal{X}$. 
What is the optimal classifier for cost-sensitive (zero-one loss) risk?

Let \( \eta(x) := P(Y = 1 \mid X = x) \) for \( x \in \mathcal{X} \).

Conditional on \( X = x \), the minimizer of conditional cost-sensitive risk

\[
\hat{y} \mapsto \mathbb{E} \left[ \ell(c)(Y, \hat{y}) \mid X = x \right]
\]

is

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

\[
\hat{y} := \begin{cases} 
+1 & \text{if } \eta(x) \cdot (1 - c) > (1 - \eta(x)) \cdot c, \\
-1 & \text{otherwise}. 
\end{cases}
\]
What is the optimal classifier for cost-sensitive (zero-one loss) risk?

Let $\eta(x) := \mathbb{P}(Y = 1 | X = x)$ for $x \in \mathcal{X}$.

Conditional on $X = x$, the minimizer of conditional cost-sensitive risk

$$\hat{y} \mapsto \mathbb{E} \left[ \ell(c)(Y, \hat{y}) \mid X = x \right]$$

$$= \eta(x) \cdot (1 - c) \cdot 1\{\hat{y} = -1\} + (1 - \eta(x)) \cdot c \cdot 1\{\hat{y} = +1\}$$

is

$$\hat{y} := \begin{cases} +1 & \text{if } \eta(x) > c, \\ -1 & \text{otherwise}. \end{cases}$$
Minimizer of cost-sensitive risk

What is the optimal classifier for cost-sensitive (zero-one loss) risk? Let \( \eta(x) := \mathbb{P}(Y = 1 \mid X = x) \) for \( x \in \mathcal{X} \).

- Conditional on \( X = x \), the minimizer of conditional cost-sensitive risk

\[
\hat{y} \mapsto \mathbb{E} \left[ \ell(c)(Y, \hat{y}) \mid X = x \right] = \eta(x) \cdot (1 - c) \cdot 1\{\hat{y} = -1\} + (1 - \eta(x)) \cdot c \cdot 1\{\hat{y} = +1\}
\]

is

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

- Therefore, the classifier based on scoring function

\[
h(x) := \eta(x) - c, \quad x \in \mathcal{X}
\]

has the smallest cost-sensitive risk \( \mathcal{R}^{(c)} \).
Minimizer of cost-sensitive risk

What is the optimal classifier for cost-sensitive (zero-one loss) risk?
Let \( \eta(x) := \mathbb{P}(Y = 1 \mid X = x) \) for \( x \in \mathcal{X} \).

- Conditional on \( X = x \), the minimizer of conditional cost-sensitive risk
  \[
  \hat{y} \mapsto \mathbb{E} \left[ \ell^{(c)}(Y, \hat{y}) \mid X = x \right] = \eta(x) \cdot (1 - c) \cdot \mathbb{1}\{\hat{y} = -1\} + (1 - \eta(x)) \cdot c \cdot \mathbb{1}\{\hat{y} = +1\}
  \]
  is
  \[
  \hat{y} := \begin{cases}
  +1 & \text{if } \eta(x) > c, \\
  -1 & \text{otherwise}.
  \end{cases}
  \]

- Therefore, the classifier based on scoring function
  \[
  h(x) := \eta(x) - c, \quad x \in \mathcal{X}
  \]
  has the smallest cost-sensitive risk \( \mathcal{R}^{(c)} \).

But where does \( c \) come from?
Common performance criteria

- **Precision**:
  \[ P(Y = +1 | h(X)) > 0 \]

- **Recall** (a.k.a. True Positive Rate):
  \[ P(h(X) > 0 | Y = +1) \]
  Same as \[ 1 - \text{False Negative Rate} \].

- **False Positive Rate**:
  \[ P(h(X) > 0 | Y = -1) \]
  Same as \[ 1 - \text{True Negative Rate} \].
Common performance criteria

- **Precision**: $\mathbb{P}(Y = +1 \mid h(X) > 0)$. 
Common performance criteria

- **Precision:**
  \[ P(Y = +1 \mid h(X) > 0). \]

- **Recall (a.k.a. True Positive Rate):**
  \[ P(h(X) > 0 \mid Y = +1). \]

  Same as 1 − **False Negative Rate**.
Common performance criteria

- **Precision**:
  \[ P(Y = +1 \mid h(X) > 0). \]

- **Recall (a.k.a. True Positive Rate)**:
  \[ P(h(X) > 0 \mid Y = +1). \]
  Same as \(1 - \text{False Negative Rate}\).

- **False Positive Rate**:
  \[ P(h(X) > 0 \mid Y = -1). \]
  Same as \(1 - \text{True Negative Rate}\).
Common performance criteria

- **Precision:**
  \[ P(Y = +1 \mid h(X) > 0). \]

- **Recall (a.k.a. True Positive Rate):**
  \[ P(h(X) > 0 \mid Y = +1). \]
  Same as \( 1 - False Negative Rate. \)

- **False Positive Rate:**
  \[ P(h(X) > 0 \mid Y = -1). \]
  Same as \( 1 - True Negative Rate. \)

- ...
Suppose you are care about *Balanced Error Rate (BER)*:

\[ BER := \frac{1}{2} \cdot \text{False Negative Rate} + \frac{1}{2} \cdot \text{False Positive Rate}. \]

Which cost-sensitive risk should you (try to) minimize?
Example: balanced error rate

Suppose you are care about *Balanced Error Rate (BER)*:

\[
\text{BER} := \frac{1}{2} \cdot \text{False Negative Rate} + \frac{1}{2} \cdot \text{False Positive Rate}.
\]

Which cost-sensitive risk should you (try to) minimize?

\[
2 \cdot \text{BER} = \underbrace{\mathbb{P}(h(X) \leq 0 \mid Y = +1)}_{\text{FNR}} + \underbrace{\mathbb{P}(h(X) > 0 \mid Y = -1)}_{\text{FPR}}
\]
Example: balanced error rate

Suppose you are care about *Balanced Error Rate (BER)*:

\[
\text{BER} := \frac{1}{2} \cdot \text{False Negative Rate} + \frac{1}{2} \cdot \text{False Positive Rate}.
\]

Which cost-sensitive risk should you (try to) minimize?

\[
2 \cdot \text{BER} = \underbrace{\mathbb{P}(h(X) \leq 0 \mid Y = +1)}_{\text{FNR}} + \underbrace{\mathbb{P}(h(X) > 0 \mid Y = -1)}_{\text{FPR}}
\]

\[
= \frac{1}{\pi} \mathbb{P}(h(X) \leq 0 \land Y = +1) + \frac{1}{1 - \pi} \mathbb{P}(h(X) > 0 \land Y = -1)
\]

where \(\pi := \mathbb{P}(Y = +1)\).
Example: balanced error rate

Suppose you are care about *Balanced Error Rate (BER)*:

\[
\text{BER} := \frac{1}{2} \cdot \text{False Negative Rate} + \frac{1}{2} \cdot \text{False Positive Rate}.
\]

Which cost-sensitive risk should you (try to) minimize?

\[
2 \cdot \text{BER} = \underbrace{\mathbb{P}(h(X) \leq 0 \mid Y = +1)}_{\text{FNR}} + \underbrace{\mathbb{P}(h(X) > 0 \mid Y = -1)}_{\text{FPR}}
\]

\[
= \frac{1}{\pi} \mathbb{P}(h(X) \leq 0 \land Y = +1) + \frac{1}{1 - \pi} \mathbb{P}(h(X) > 0 \land Y = -1)
\]

where \( \pi := \mathbb{P}(Y = +1) \). So use \( R^{(c)} \) with

\[
c := \frac{1}{1 - \pi} = \pi = \mathbb{P}(Y = +1)
\]

(which you can estimate).
Even more general setting

**Distribution over importance-weighted examples:**
Every example comes with example-specific weight:

$$(X, Y, W) \sim P$$

where $W$ is the (non-negative) *importance weight* of the example.
Even more general setting

Distribution over importance-weighted examples:
Every example comes with example-specific weight:

\[(X, Y, W) \sim P\]

where \(W\) is the (non-negative) \textit{importance weight} of the example.

Importance-weighted \(\ell\)-risk of \(h\) is

\[
\mathcal{R}(h) := \mathbb{E} \left[ W \cdot \ell(Yh(X)) \right].
\]
Even more general setting

**Distribution over importance-weighted examples:**
Every example comes with example-specific weight:

\[(X, Y, W) \sim P\]

where \(W\) is the (non-negative) *importance weight* of the example.

Importance-weighted \(\ell\)-risk of \(h\) is

\[\mathcal{R}(h) := \mathbb{E} \left[ W \cdot \ell(Yh(X)) \right].\]

(This comes up in Boosting, online decision-making, etc.)
3. Conditional probability estimation
Eliciting conditional probabilities

Having the conditional probability function $\eta$ permits probabilistic reasoning. Can we estimate $\eta$ by minimizing a particular (empirical) risk?
Eliciting conditional probabilities

Having the conditional probability function $\eta$ permits probabilistic reasoning. Can we estimate $\eta$ by minimizing a particular (empirical) risk?

- **Squared loss**: $\ell_{sq}(z) = (1 - z)^2$

Using calculus, can see that this is minimized by $h$ s.t. $h(x) = 2\eta(x) - 1$, i.e., $\eta(x) = 1 + h(x)^2$.

Recipe using plug-in principle:

1. Find scoring function $\hat{h} : X \rightarrow \mathbb{R}$ to (approximately) minimize (empirical) squared loss risk.
2. Construct estimate $\hat{\eta}$ via $\hat{\eta}(x) := 1 + h(x)^2$, $x \in X$.

(May want to clip output to range $[0, 1]$. )
Eliciting conditional probabilities

Having the conditional probability function $\eta$ permits probabilistic reasoning. Can we estimate $\eta$ by minimizing a particular (empirical) risk?

- **Squared loss**: $\ell_{sq}(z) = (1 - z)^2$

\[
\mathbb{E}[\ell_{sq}(Yh(x)) \mid X = x] = \eta(x) \cdot (1 - h(x))^2 + (1 - \eta(x)) \cdot (1 + h(x))^2.
\]
Eliciting conditional probabilities

Having the conditional probability function \( \eta \) permits probabilistic reasoning. Can we estimate \( \eta \) by minimizing a particular (empirical) risk?

- **Squared loss**: \( \ell_{sq}(z) = (1 - z)^2 \)

\[
\mathbb{E}[\ell_{sq}(Yh(x)) \mid X = x] = \eta(x) \cdot (1 - h(x))^2 + (1 - \eta(x)) \cdot (1 + h(x))^2.
\]

Using calculus, can see that this is minimized by \( h \) s.t. \( h(x) = 2\eta(x) - 1 \), i.e., \( \eta(x) = \frac{1 + h(x)}{2} \).
Eliciting conditional probabilities

Having the conditional probability function $\eta$ permits probabilistic reasoning. Can we estimate $\eta$ by minimizing a particular (empirical) risk?

▶ **Squared loss:** $\ell_{sq}(z) = (1 - z)^2$

$$
\mathbb{E}[\ell_{sq}(Yh(x)) \mid X = x] = \eta(x) \cdot (1 - h(x))^2 + (1 - \eta(x)) \cdot (1 + h(x))^2.
$$

Using calculus, can see that this is minimized by $h$ s.t. $h(x) = 2\eta(x) - 1$, i.e., $\eta(x) = \frac{1 + h(x)}{2}$.

**Recipe using plug-in principle:**

1. Find scoring function $\hat{h} : \mathcal{X} \to \mathbb{R}$ to (approximately) minimize (empirical) squared loss risk.
2. Construct estimate $\hat{\eta}$ via

$$
\hat{\eta}(x) := \frac{1 + \hat{h}(x)}{2}, \quad x \in \mathcal{X}.
$$

(May want to clip output to range $[0, 1]$.)
Eliciting conditional probabilities (again)

Logistic loss: 
\[ \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) \]

This is minimized by 
\[ h \text{ s.t. } \ell_{\text{logistic}}(h(x)) = \eta(x) \]

Recipe using plug-in principle:
1. Find scoring function \( \hat{h} : X \rightarrow \mathbb{R} \) to (approximately) minimize (empirical) logistic loss risk.
2. Construct estimate \( \hat{\eta} \) via \( \hat{\eta}(x) := \text{logistic}(h(x)) \), \( x \in X \).

This works for many other losses . . .
Eliciting conditional probabilities (again)

- **Logistic loss:** \( \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) \)
Eliciting conditional probabilities (again)

- **Logistic loss**: \( \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) \)

\[
\mathbb{E}[\ell_{\text{logistic}}(Y h(x)) \mid X = x] = \eta(x) \cdot \log_2(1+e^{-h(x)}) + (1-\eta(x)) \cdot \log_2(1+e^{h(x)}).
\]
Eliciting conditional probabilities (again)

- **Logistic loss**: \( \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) \)

\[
E[\ell_{\text{logistic}}(Y h(x)) \mid X = x] = \eta(x) \cdot \log_2(1 + e^{-h(x)}) + (1 - \eta(x)) \cdot \log_2(1 + e^{h(x)}).
\]

This is minimized by \( h \) s.t. \( \text{logistic}(h(x)) = \eta(x) \).
Eliciting conditional probabilities (again)

- **Logistic loss**: $\ell_{\text{logistic}}(z) = \log_2(1 + e^{-z})$

$$\mathbb{E}[\ell_{\text{logistic}}(Yh(x)) \mid X = x] = \eta(x) \cdot \log_2(1 + e^{-h(x)}) + (1 - \eta(x)) \cdot \log_2(1 + e^{h(x)})$$

This is minimized by $h \text{ s.t. } \text{logistic}(h(x)) = \eta(x)$.

**Recipe using plug-in principle:**

1. Find scoring function $\hat{h} : \mathcal{X} \to \mathbb{R}$ to (approximately) minimize (empirical) logistic loss risk.
2. Construct estimate $\hat{\eta}$ via

$$\hat{\eta}(x) := \text{logistic}(h(x)), \quad x \in \mathcal{X}.$$
Eliciting conditional probabilities (again)

- **Logistic loss:** \( \ell_{\text{logistic}}(z) = \log_2(1 + e^{-z}) \)

\[
E[\ell_{\text{logistic}}(Yh(x)) \mid X = x] = \eta(x) \cdot \log_2(1 + e^{-h(x)}) + (1 - \eta(x)) \cdot \log_2(1 + e^{h(x)}).
\]

This is minimized by \( h \) s.t. \( \text{logistic}(h(x)) = \eta(x) \).

**Recipe using plug-in principle:**

1. Find scoring function \( \hat{h} : \mathcal{X} \rightarrow \mathbb{R} \) to (approximately) minimize (empirical) logistic loss risk.
2. Construct estimate \( \hat{\eta} \) via

\[
\hat{\eta}(x) := \text{logistic}(h(x)), \quad x \in \mathcal{X}.
\]

- This works for many other losses . . .
Deficiency of hinge loss

But not **hinge loss**: \( \ell_{\text{hinge}}(z) = \max\{0, 1 - z\} \).
Deficiency of hinge loss

But not hinge loss: $\ell_{\text{hinge}}(z) = \max\{0, 1 - z\}$.

$$\mathbb{E}[\ell_{\text{hinge}}(Y h(x)) \mid X = x]$$

is minimized by $h$ s.t. $h(x) = \text{sign}(2\eta(x) - 1)$.
Deficiency of hinge loss

But not hinge loss: \( \ell_{\text{hinge}}(z) = \max\{0, 1 - z\} \).

\[
\mathbb{E}[\ell_{\text{hinge}}(Y h(x)) \mid X = x]
\]
is minimized by \( h \) s.t. \( h(x) = \text{sign}(2\eta(x) - 1) \).

Cannot recover \( \eta(x) \) from the minimizing \( h(x) \).
**Caveat**: In general,

$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\ell_{sq}(YX^Tw)] \geq \min_{h: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E}[\ell_{sq}(Yh(X))].$$

Optimal scoring function $x \mapsto 2\eta(x) - 1$ is not necessarily well-approximated by a linear function.
**Caveat**: In general,

\[
\min_{w \in \mathbb{R}^d} \mathbb{E}[\ell_{sq}(Y X^T w)] \geq \min_{h: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E}[\ell_{sq}(Y h(X))].
\]

Optimal scoring function \( x \mapsto 2\eta(x) - 1 \) is not necessarily well-approximated by a linear function. (We’ll see an example of this later.)
Caveat: In general,

$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\ell_{sq}(Y X^T w)] \geq \min_{h: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E}[\ell_{sq}(Y h(X))].$$

Optimal scoring function $x \mapsto 2\eta(x) - 1$ is not necessarily well-approximated by a linear function. (We’ll see an example of this later.)

Common remedies: enhance the feature space via feature expansion, or use more flexible models (e.g., decision trees, neural nets).
Square loss vs. logistic loss

Do square and logistic loss care about conditional probabilities the same way?
Square loss vs. logistic loss

Do square and logistic loss care about conditional probabilities the same way?

- **Square loss** in terms of \( p(x) := \frac{1 + h(x)}{2} \):

\[
\ell_{sq}(yh(x)) = 4 \left( \mathbbm{1}\{y = 1\} \cdot (1 - p(x))^2 + \mathbbm{1}\{y = -1\} \cdot (0 - p(x))^2 \right)
\]
Square loss vs. logistic loss

Do square and logistic loss care about conditional probabilities the same way?

- **Square loss** in terms of \( p(x) := \frac{1 + h(x)}{2} \):

\[
\ell_{sq}(yh(x)) = 4 \left( 1\{y = 1\} \cdot (1 - p(x))^2 + 1\{y = -1\} \cdot (0 - p(x))^2 \right)
\]

\[
\propto \int_{0}^{1} \ell_{0/1}^{(c)} \left( y(p(x) - c) \right) \cdot dc.
\]
Square loss vs. logistic loss

Do square and logistic loss care about conditional probabilities the same way?

- **Square loss** in terms of \( p(x) := \frac{1 + h(x)}{2} \):

  \[
  \ell_{\text{sq}}(yh(x)) = 4 \left( \mathbbm{1}\{y = 1\} \cdot (1 - p(x))^2 + \mathbbm{1}\{y = -1\} \cdot (0 - p(x))^2 \right)
  \propto \int_0^1 \ell_{0/1}^c (y(p(x) - c)) \cdot dc.
  \]

- **Logistic loss** in terms of \( p(x) := \text{logistic}(h(x)) \):

  \[
  \ell_{\text{logistic}}(yh(x)) = \mathbbm{1}\{y = 1\} \cdot \log_2 \frac{1}{p(x)} + \mathbbm{1}\{y = -1\} \cdot \log_2 \frac{1}{1 - p(x)}
  \]

Logistic loss cares about accuracy of probabilities near 0 and 1 more than it cares about probabilities near \( \frac{1}{2} \).
Square loss vs. logistic loss

Do square and logistic loss care about conditional probabilities the same way?

- **Square loss** in terms of \( p(x) := \frac{1+h(x)}{2} \):

  \[
  \ell_{sq}(yh(x)) = 4 \left( \mathbb{1}\{y = 1\} \cdot (1 - p(x))^2 + \mathbb{1}\{y = -1\} \cdot (0 - p(x))^2 \right)
  \]
  \[
  \propto \int_{0}^{1} \ell_{0/1}^{(c)} \left( y(p(x) - c) \right) \cdot dc.
  \]

- **Logistic loss** in terms of \( p(x) := \text{logistic}(h(x)) \):

  \[
  \ell_{\text{logistic}}(yh(x)) = \mathbb{1}\{y = 1\} \cdot \log_2 \frac{1}{p(x)} + \mathbb{1}\{y = -1\} \cdot \log_2 \frac{1}{1 - p(x)}
  \]
  \[
  \propto \int_{0}^{1} \ell_{0/1}^{(c)} \left( Y(p(x) - c) \right) \cdot \frac{dc}{c(1 - c)}.
  \]
Do square and logistic loss care about conditional probabilities the same way?

- **Square loss** in terms of $p(x) := \frac{1 + h(x)}{2}$:

$$
\ell_{\text{sq}}(y h(x)) = 4 \left( \mathbb{1}\{y = 1\} \cdot (1 - p(x))^2 + \mathbb{1}\{y = -1\} \cdot (0 - p(x))^2 \right)
\propto \int_0^1 \ell_{0/1}^{(c)}(y(p(x) - c)) \cdot dc.
$$

- **Logistic loss** in terms of $p(x) := \text{logistic}(h(x))$:

$$
\ell_{\text{logistic}}(y h(x)) = \mathbb{1}\{y = 1\} \cdot \log_2 \frac{1}{p(x)} + \mathbb{1}\{y = -1\} \cdot \log_2 \frac{1}{1 - p(x)}
\propto \int_0^1 \ell_{0/1}^{(c)}(Y(p(x) - c)) \cdot \frac{dc}{c(1 - c)}.
$$

Logistic loss cares about accuracy of probabilities near 0 and 1 more than it cares about probabilities near 1/2.
Calibration

Say $\hat{\eta}$ is *calibrated* if

$$\mathbb{P}(Y = 1 \mid \hat{\eta}(X) = p) = p, \quad p \in [0, 1].$$

E.g., It rains on half of the days on which you say "50% chance of rain".

Note: This is achieved by a constant function:

$$\hat{\eta}(x) = \mathbb{P}(Y = 1), \quad x \in X.$$
Say \( \hat{\eta} \) is \textit{calibrated} if

\[
\mathbb{P}(Y = 1 \mid \hat{\eta}(X) = p) = p, \quad p \in [0, 1].
\]

E.g., It rains on half of the days on which you say “50% chance of rain”.

\[\text{Calibration}\]
Say \( \hat{\eta} \) is \textit{calibrated} if

\[
\mathbb{P}(Y = 1 \mid \hat{\eta}(X) = p) = p, \quad p \in [0, 1].
\]

E.g., It rains on half of the days on which you say “50% chance of rain”.

\textbf{Note:} This is achieved by a constant function:

\[
\hat{\eta}(x) = \mathbb{P}(Y = 1), \quad x \in \mathcal{X}.
\]

So calibration is weaker than approximating the conditional probability function.
Achieving empirical calibration (isotonic regression)

Let $\hat{\eta}: \mathcal{X} \rightarrow [0, 1]$ be an estimate of the conditional probability function.

**Goal:** Find non-decreasing $\hat{g}: \mathbb{R} \rightarrow \mathbb{R}$ such that $\hat{g}(\hat{\eta}(\cdot))$ is calibrated w.r.t. empirical distribution $P_n$ on $((x_i, y_i))_{i=1}^n$. 

In this figure, we assume $\hat{\eta}(x_1) < \cdots < \hat{\eta}(x_n)$. 

Achieving empirical calibration (isotonic regression)

Let $\hat{\eta}: \mathcal{X} \to [0, 1]$ be an estimate of the conditional probability function.

**Goal:** Find non-decreasing $\hat{g}: \mathbb{R} \to \mathbb{R}$ such that $\hat{g}(\hat{\eta}(\cdot))$ is calibrated w.r.t. empirical distribution $P_n$ on $((x_i, y_i))_{i=1}^n$.

This is achieved by solving

$$
\min_{u_1, \ldots, u_n \in \mathbb{R}} \sum_{i=1}^{n} (u_i - 1\{y_i = 1\})^2
$$

subject to $u_i \leq u_j$ for all $i, j$ s.t. $\hat{\eta}(x_i) \leq \hat{\eta}(x_j)$

and letting $\hat{g}$ linearly interpolate $((\hat{\eta}(x_i), u_i))_{i=1}^n$.
Achieving empirical calibration (isotonic regression)

Let \( \hat{\eta} : \mathcal{X} \rightarrow [0, 1] \) be an estimate of the conditional probability function.

**Goal:** Find non-decreasing \( \hat{g} : \mathbb{R} \rightarrow \mathbb{R} \) such that \( \hat{g}(\hat{\eta}(\cdot)) \) is calibrated w.r.t. empirical distribution \( P_n \) on \( ((x_i, y_i))_{i=1}^n \).

This is achieved by solving

\[
\min_{u_1, \ldots, u_n \in \mathbb{R}} \sum_{i=1}^n (u_i - 1\{y_i = 1\})^2 \\
\text{s.t.} \quad u_i \leq u_j \quad \text{for all } i, j \text{ s.t. } \hat{\eta}(x_i) \leq \hat{\eta}(x_j)
\]

and letting \( \hat{g} \) linearly interpolate \( ((\hat{\eta}(x_i), u_i))_{i=1}^n \).

(In this figure, we assume \( \hat{\eta}(x_1) < \cdots < \hat{\eta}(x_n) \).)
4. Multiclass classification
Multiclass classification

Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (*multiclass classification*).
Multiclass classification

Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (*multiclass classification*).

- **Decision trees**: already formulated for multiclass.
Multiclass classification

Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (*multiclass classification*).

- **Decision trees**: already formulated for multiclass.
- **Logistic regression**: generalizes to *multinomial logistic regression*, where

$$
P(Y = k \mid X = x) = \frac{e^{x^T \beta_k}}{\sum_{k'=1}^K e^{x^T \beta_{k'}}}, \quad x \in \mathbb{R}^d.
$$

MLE here is similar to MLE in binary case.

- **Perceptron, SVM**: many extensions for multiclass.
- **Neural networks**: often formulated like multinomial logistic regression.

Can we generically reduce multiclass classification to binary classification?
Multiclass classification

Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (*multiclass classification*).

- **Decision trees**: already formulated for multiclass.
- **Logistic regression**: generalizes to *multinomial logistic regression*, where

  \[ \mathbb{P}(Y = k \mid X = x) = \frac{e^{x^T \beta_k}}{\sum_{k' = 1}^{K} e^{x^T \beta_{k'}}}, \quad x \in \mathbb{R}^d. \]

  MLE here is similar to MLE in binary case.

- **Perceptron, SVM**: many extensions for multiclass.
Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (*multiclass classification*).

- **Decision trees**: already formulated for multiclass.
- **Logistic regression**: generalizes to *multinomial logistic regression*, where

$$
\mathbb{P}(Y = k \mid X = x) = \frac{e^{x^T \beta_k}}{\sum_{k'=1}^{K} e^{x^T \beta_{k'}}}, \quad x \in \mathbb{R}^d.
$$

MLE here is similar to MLE in binary case.

- **Perceptron, SVM**: many extensions for multiclass.
- **Neural networks**: often formulated like multinomial logistic regression.
Often have more than two classes $\mathcal{Y} = \{1, \ldots, K\}$ (multiclass classification).

- **Decision trees**: already formulated for multiclass.
- **Logistic regression**: generalizes to *multinomial logistic regression*, where

$$
\Pr(Y = k \mid X = x) = \frac{e^{x^T \beta_k}}{\sum_{k'=1}^K e^{x^T \beta_{k'}}}, \quad x \in \mathbb{R}^d.
$$

MLE here is similar to MLE in binary case.

- **Perceptron, SVM**: many extensions for multiclass.
- **Neural networks**: often formulated like multinominal logistic regression.

Can we generically reduce multiclass classification to binary classification?
One-against-all (of the rest) reduction

### One-against-all training

**input** Labeled examples \( \left( (x_i, y_i) \right)_{i=1}^{n} \) from \( \mathcal{X} \times \{1, \ldots, K\} \); learning algorithm \( \mathcal{A} \) that takes data from \( \mathcal{X} \times \{-1, 1\} \) and returns a scoring function \( h : \mathcal{X} \rightarrow \mathbb{R} \).

**output** Scoring functions \( \hat{h}_1, \ldots, \hat{h}_K : \mathcal{X} \rightarrow \mathbb{R} \).

1. **for** \( k = 1, \ldots, K \) **do**
   2. Create data set \( S_k = \left( (x_i, y_i^{(k)}) \right)_{i=1}^{n} \) where \( y_i^{(k)} = \begin{cases} +1 & \text{if } y_i = k, \\ -1 & \text{otherwise} \end{cases} \).
   3. Run \( \mathcal{A} \) on \( S_k \) to obtain scoring function \( h_k : \mathcal{X} \rightarrow \mathbb{R} \).
4. **end for**
5. **return** Scoring functions \( \hat{h}_1, \ldots, \hat{h}_K \).
One-against-all (of the rest) reduction

**One-against-all training**

**input** Labeled examples $((x_i, y_i))_{i=1}^n$ from $\mathcal{X} \times \{1, \ldots, K\}$; learning algorithm $\mathcal{A}$ that takes data from $\mathcal{X} \times \{\pm 1\}$ and returns a scoring function $h: \mathcal{X} \rightarrow \mathbb{R}$.

**output** Scoring functions $\hat{h}_1, \ldots, \hat{h}_K: \mathcal{X} \rightarrow \mathbb{R}$.

1. for $k = 1, \ldots, K$ do
2. Create data set $S_k = ((x_i, y_i^{(k)}))_{i=1}^n$ where $y_i^{(k)} = \begin{cases} +1 & \text{if } y_i = k, \\ -1 & \text{otherwise.} \end{cases}$
3. Run $\mathcal{A}$ on $S_k$ to obtain scoring function $h_k: \mathcal{X} \rightarrow \mathbb{R}$.
4. end for
5. return Scoring functions $\hat{h}_1, \ldots, \hat{h}_K$.

**One-against-all prediction**

**input** Scoring functions $\hat{h}_1, \ldots, \hat{h}_K: \mathcal{X} \rightarrow \mathbb{R}$; new point $x \in \mathcal{X}$.

**output** Predicted label $\hat{y} \in \{1, \ldots, K\}$.

1. return $\hat{y} \in \arg \max_{k \in \{1, \ldots, K\}} \hat{h}_k(x)$. 
Main idea: Suppose every $\hat{h}_k$ corresponds to good estimate $\hat{\eta}_k$ of conditional probability function

$$\hat{\eta}_k(x) \approx \Pr(Y = k \mid X = x)$$

on average. (E.g., $\hat{\eta}_k(x) = \frac{1 + \hat{h}_k(x)}{2}$ for all $k = 1, \ldots, K$.)

When does one-against-all work well?
Main idea: Suppose every $\hat{h}_k$ corresponds to good estimate $\hat{\eta}_k$ of conditional probability function

$$\hat{\eta}_k(x) \approx \mathbb{P}(Y = k \mid X = x)$$

on average. (E.g., $\hat{\eta}_k(x) = \frac{1+\hat{h}_k(x)}{2}$ for all $k = 1, \ldots, K$.)

$\Rightarrow$ Behavior of one-against-all classifier $\hat{f}$ similar to Bayes optimal classifier!
Suppose $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{1, 2, 3\}$, and $P_{Y|X=x}$ is piece-wise constant:

$P(Y = 1 | X = x) = 1$

$P(Y = 2 | X = x) = 1$

$P(Y = 3 | X = x) = 1$

Try to fit $P(Y = k | X = x)$ using affine functions of $x$.

The arg max $k$ is never 2, so never predicts label 2.

How to fix this?
Suppose $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{1, 2, 3\}$, and $P_{Y|X=x}$ is piece-wise constant:

- $P(Y = 1 | X = x) = 1$
- $P(Y = 2 | X = x) = 1$
- $P(Y = 3 | X = x) = 1$

Try to fit $P(Y = k | X = x)$ using affine functions of $x$. 
Suppose $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{1, 2, 3\}$, and $P_{Y|X=x}$ is piece-wise constant:

$$
P(Y = 1 | X = x) = 1 \quad \quad P(Y = 2 | X = x) = 1 \quad \quad P(Y = 3 | X = x) = 1
$$

Try to fit $P(Y = k | X = x)$ using affine functions of $x$.

The $\text{arg max}_k$ is never 2, so never predicts label 2.
Suppose $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{1, 2, 3\}$, and $P_{Y \mid X=x}$ is piece-wise constant:

\[
P(Y = 1 \mid X = x) = 1 \\
P(Y = 2 \mid X = x) = 1 \\
P(Y = 3 \mid X = x) = 1
\]

Try to fit $P(Y = k \mid X = x)$ using affine functions of $x$.

The $\arg \max_k$ is never 2, so never predicts label 2. How to fix this?
Key takeaways

1. Scoring functions and thresholds; alternative performance criteria.
2. Eliciting conditional probabilities with loss functions.
3. Calibration property.
4. One-against-all approach to multiclass classification; masking.