The Bayes Optimal Classifier

\[ f^{(BO)}(x) = \arg\max_y D(x, y) \]
The Bayes Optimal Classifier

\[ f^{(BO)}(x) = \arg\max_y D(x, y) \]

**Theorem:** The Bayes optimal classifier achieves minimal zero/one error \((\ell(y, \hat{y}) = [y \neq \hat{y}])\) of any deterministic classifier.
Consider (deterministic) $f'$ that claims to be better than $f^{(BO)}$ and $x$ such that $f^{(BO)}(x) \neq f'(x)$. 
Proof

Consider (deterministic) $f'$ that claims to be better than $f^{(BO)}$ and $x$ such that $f^{(BO)}(x) \neq f'(x)$.

Probability that $f'$ makes an error on this input: $\left(\sum_y D(x, y)\right) - D(x, f'(x))$. 
Proof

Consider (deterministic) $f'$ that claims to be better than $f^{(BO)}$ and $x$ such that $f^{(BO)}(x) \neq f'(x)$.

Probability that $f'$ makes an error on this input: $\left( \sum_y D(x, y) \right) - D(x, f'(x))$.

Probability that $f^{(BO)}$ makes an error on this input: $\left( \sum_y D(x, y) \right) - D(x, f^{(BO)}(x))$. 
Proof

Consider (deterministic) $f'$ that claims to be better than $f^{(BO)}$ and $x$ such that $f^{(BO)}(x) \neq f'(x)$.

Probability that $f'$ makes an error on this input: 

$$\left(\sum_y \mathcal{D}(x, y)\right) - \mathcal{D}(x, f'(x)).$$

Probability that $f^{(BO)}$ makes an error on this input: 

$$\left(\sum_y \mathcal{D}(x, y)\right) - \mathcal{D}(x, f^{(BO)}(x)).$$

By definition,

$$\mathcal{D}(x, f^{(BO)}(x)) = \max_y \mathcal{D}(x, y) \geq \mathcal{D}(x, f'(x))$$

$$\Rightarrow \left(\sum_y \mathcal{D}(x, y)\right) - \mathcal{D}(x, f^{(BO)}(x)) \leq \left(\sum_y \mathcal{D}(x, y)\right) - \mathcal{D}(x, f'(x))$$
Proof

Consider (deterministic) $f'$ that claims to be better than $f^{(BO)}$ and $x$ such that $f^{(BO)}(x) \neq f'(x)$.

Probability that $f'$ makes an error on this input: $\left(\sum_y D(x, y)\right) - D(x, f'(x))$.

Probability that $f^{(BO)}$ makes an error on this input: $\left(\sum_y D(x, y)\right) - D(x, f^{(BO)}(x))$.

By definition,

$$D(x, f^{(BO)}(x)) = \max_y D(x, y) \geq D(x, f'(x))$$

$$\Rightarrow \left(\sum_y D(x, y)\right) - D(x, f^{(BO)}(x)) \leq \left(\sum_y D(x, y)\right) - D(x, f'(x))$$

This must hold for all $x$. Hence $f'$ is no better than $f^{(BO)}$. 
One Limit of Learning

You cannot do better than \( \epsilon(f^{BO}) \).
Unavoidable Error

- Noise in the features (we don’t want to “fit” the noise!)
- Insufficient information in the available features (e.g., incomplete data)
- No single correct label (e.g., inconsistencies in the data-generating process)

These have nothing to do with your choice of learning algorithm.
An Exercise
Following Daume (2017), chapter 2.

Class A

Class B
An Exercise
Following Daume (2017), chapter 2.

Test
Inductive Bias

Just as you had a tendency to focus on a certain type of function $f$, machine learning algorithms correspond to classes of functions ($\mathcal{F}$) and preferences within the class.
Inductive Bias

Just as *you* had a tendency to focus on a certain type of function \( f \), machine learning algorithms correspond to classes of functions (\( F \)) and preferences within the class.

E.g., shallow decision trees: “use a small number of features.”
The cardinal rule of machine learning: **Don’t touch your test data.**

If you follow that rule, this recipe will give you accurate information:

1. Split data into training, development, and test sets.
2. For different hyperparameter settings:
   2.1 Train on the training data using those hyperparameter values.
   2.2 Evaluate loss on development data.
3. Choose the hyperparameter setting whose model achieved the lowest development data loss.
   Optionally, retrain on the training and development data together.
4. Evaluate that model on test data.
Design Process for ML Applications

1. real world goal
2. mechanism
3. learning problem
4. data collection
5. collected data
6. data representation
7. select model family
8. select training/dev. data
9. train and select hyperparameters
10. make predictions on test set
11. evaluate error
12. deploy

example

increase revenue
show better ads
will a user who queries \( q \) click ad \( a \) ?
interaction with existing system
query \( q, a, \pm \)click
(\( q \) word, \( a \) word) pairs
decision trees up to 20
September
single decision tree
October
zero-one loss (\( \pm \)click)
$?
## Features

Data derived from [https://archive.ics.uci.edu/ml/datasets/Auto+MPG](https://archive.ics.uci.edu/ml/datasets/Auto+MPG)

<table>
<thead>
<tr>
<th>mpg</th>
<th>cylinders</th>
<th>displacement</th>
<th>horsepower</th>
<th>weight</th>
<th>acceleration</th>
<th>year</th>
<th>origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.0</td>
<td>8</td>
<td>307.0</td>
<td>130.0</td>
<td>3504.</td>
<td>12.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>15.0</td>
<td>8</td>
<td>350.0</td>
<td>165.0</td>
<td>3693.</td>
<td>11.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>18.0</td>
<td>8</td>
<td>318.0</td>
<td>150.0</td>
<td>3436.</td>
<td>11.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>16.0</td>
<td>8</td>
<td>304.0</td>
<td>150.0</td>
<td>3433.</td>
<td>12.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>17.0</td>
<td>8</td>
<td>302.0</td>
<td>140.0</td>
<td>3449.</td>
<td>10.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>15.0</td>
<td>8</td>
<td>429.0</td>
<td>198.0</td>
<td>4341.</td>
<td>10.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>14.0</td>
<td>8</td>
<td>454.0</td>
<td>220.0</td>
<td>4354.</td>
<td>9.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>14.0</td>
<td>8</td>
<td>440.0</td>
<td>215.0</td>
<td>4312.</td>
<td>8.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>14.0</td>
<td>8</td>
<td>455.0</td>
<td>225.0</td>
<td>4425.</td>
<td>10.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>15.0</td>
<td>8</td>
<td>390.0</td>
<td>190.0</td>
<td>3850.</td>
<td>8.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>15.0</td>
<td>8</td>
<td>383.0</td>
<td>170.0</td>
<td>3563.</td>
<td>10.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>14.0</td>
<td>8</td>
<td>340.0</td>
<td>160.0</td>
<td>3609.</td>
<td>8.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>15.0</td>
<td>8</td>
<td>400.0</td>
<td>150.0</td>
<td>3761.</td>
<td>9.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>14.0</td>
<td>8</td>
<td>455.0</td>
<td>225.0</td>
<td>3086.</td>
<td>10.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>24.0</td>
<td>4</td>
<td>113.0</td>
<td>95.0</td>
<td>2372.</td>
<td>15.0</td>
<td>70</td>
<td>3</td>
</tr>
<tr>
<td>22.0</td>
<td>6</td>
<td>198.0</td>
<td>95.0</td>
<td>2833.</td>
<td>15.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>18.0</td>
<td>6</td>
<td>199.0</td>
<td>97.0</td>
<td>2774.</td>
<td>15.5</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>21.0</td>
<td>6</td>
<td>200.0</td>
<td>85.0</td>
<td>2587.</td>
<td>16.0</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>27.0</td>
<td>4</td>
<td>97.0</td>
<td>88.0</td>
<td>2130.</td>
<td>14.5</td>
<td>70</td>
<td>3</td>
</tr>
<tr>
<td>26.0</td>
<td>4</td>
<td>97.0</td>
<td>46.0</td>
<td>1835.</td>
<td>20.5</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>25.0</td>
<td>4</td>
<td>110.0</td>
<td>87.0</td>
<td>2672.</td>
<td>17.5</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>24.0</td>
<td>4</td>
<td>107.0</td>
<td>90.0</td>
<td>2430.</td>
<td>14.5</td>
<td>70</td>
<td>2</td>
</tr>
</tbody>
</table>

All features are represented as $\mathbb{R}$ values.

Side note: could convert discrete origin feature into three binary features as follows:

- 1/america $\rightarrow (1, 0, 0)$
- 2/europe $\rightarrow (0, 1, 0)$
- 3/asia $\rightarrow (0, 0, 1)$

The “1–2–3” values suggest ordinality, which is misleading.
Instance \( x \) Becomes Vector \( \mathbf{x} \)

First example in the data, “Chevrolet Chevelle Malibu,” becomes:

\[
[8, 307.0, 130.0, 3504, 12.0, 70, 1, 0, 0]
\]

“Buick Skylark 320” becomes:

\[
[8, 350.0, 165.0, 3693, 11.5, 70, 1, 0, 0]
\]
Euclidean Distance

General formula for the Euclidean distance between two $d$-length vectors:

$$dist(x, x') = \sqrt{\sum_{j=1}^{d} (x[j] - x'[j])^2}$$

$$= \| x - x' \|_2$$
Euclidean Distance

General formula for the Euclidean distance between two $d$-length vectors:

$$\text{dist}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1}^{d} (x[j] - x'[j])^2} = \| \mathbf{x} - \mathbf{x}' \|_2$$

The distance between the Chevrolet Chevelle Malibu and the Buick Skylark 320:

$$\sqrt{(8 - 8)^2 + (307 - 350)^2 + (130 - 165)^2 + (3504 - 3693)^2}$$
$$+ (12 - 11.5)^2 + (70 - 70)^2 + (1 - 1)^2 + (0 - 0)^2 + (0 - 0)^2$$
$$= \sqrt{1849 + 1225 + 35721 + 0.25}$$
$$= \sqrt{1849 + 1225 + 35721 + 0.25}$$
$$\approx 196.965$$