Classification

- Input: Data $\mathcal{D} = \{(x_t, y_t) \mid t = 1, \dots, N\}$.
- Observations: $x_t \in \mathcal{X}$, with $\mathcal{X}$ arbitrary.
- Labels: $y_t \in \mathcal{Y} = \{1, \dots, K\}$.

Classification

- Input: Data $\mathcal{D} = \{(x_t, y_t) \mid t = 1, \ldots, N\}$.
- Observations: $x_t \in \mathcal{X}$, with $\mathcal{X}$ arbitrary.
- Labels: $y_t \in \mathcal{Y} = \{1, \ldots, K\}$.

Example 1

Age	Sex	Smoking	Cancer
27	F	Yes	0
44	M	No	0
55	F	Yes	0
60	F	No	0
30	M	Yes	0
41	M	Yes	1
47	F	No	0
62	F	Yes	0
64	M	No	1

Classification

- Input: Data $\mathcal{D} = \{(x_t, y_t) \mid t = 1, \ldots, N\}$.
- Observations: $x_t \in \mathcal{X}$, with $\mathcal{X}$ arbitrary.
- Labels: $y_t \in \mathcal{Y} = \{1, \ldots, K\}$.

The input $\mathcal{X}$ is composed of features/attributes $\mathcal{A}_i$

$\mathcal{X} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_P$, with $\mathcal{A}_i$:

- Boolean, i.e. $\mathcal{A}_i = \{0, 1\}$.
- Categorical $\mathcal{A}_i = \{\mathrm{cat}, \mathrm{dog}\}$
- Real, i.e. $\mathcal{A}_i = \mathbb{R}$.

Example 1

Age	Sex	Smoking	Cancer
27	F	Yes	0
44	M	No	0
55	F	Yes	0
60	F	No	0
30	M	Yes	0
41	M	Yes	1
47	F	No	0
62	F	Yes	0
64	M	No	1

Classification

- Input: Data $\mathcal{D} = \{(x_t, y_t) \mid t = 1, \ldots, N\}$.
- Observations: $x_t \in \mathcal{X}$, with $\mathcal{X}$ arbitrary.
- Labels: $y_t \in \mathcal{Y} = \{1, \ldots, K\}$.

The input $\mathcal{X}$ is composed of features/attributes $\mathcal{A}_i$

$\mathcal{X} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_P$, with $\mathcal{A}_i$:

- Boolean, i.e. $\mathcal{A}_i = \{0, 1\}$.
- Categorical $\mathcal{A}_i = \{\mathtt{cat}, \mathtt{dog}\}$
- Real, i.e. $\mathcal{A}_i = \mathbb{R}$.

Example 1

Age	Sex	Smoking	Cancer
27	F	Yes	0
44	M	No	0
55	F	Yes	0
60	F	No	0
30	M	Yes	0
41	M	Yes	1
47	F	No	0
62	F	Yes	0
64	M	No	1

We want to find a relation between the observed attributes and the variable we want to predict.

A simple classification rule

```
if Smoking then
    if Male then
        if Age > 40 then
            Cancer
        else
            Healthy
        end if
    else
        Healthy
    end if
else
    if Age > 60 and Male then
        Cancer
    else
        Healthy
    end if
end if
```

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Probabilities as proportions of $\mathcal{D}$

$$\hat{\mathbb{P}}(y_t = i) = \frac{|\{y_t = i \mid t \in 1, \ldots, N\}|}{N}$$

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Probabilities as proportions of $\mathcal{D}$

$$\hat{\mathbb{P}}(y_t = i) = \frac{|\{y_t = i \mid t \in 1, \ldots, N\}|}{N}$$

$$\hat{\mathbb{P}}(y_t = i \mid x_t^{\text{sex}} = j) = \frac{|\{y_t = i \wedge x_t^{\text{sex}} = j \mid t \in 1, \ldots, N\}|}{|\{x_t^{\text{sex}} = j \mid t \in 1, \ldots, N\}|}$$

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Probabilities as proportions of $\mathcal{D}$

$$\hat{\mathbb{P}}(y_t = i) = \frac{|\{y_t = i \mid t \in 1, \ldots, N\}|}{N}$$

$$\hat{\mathbb{P}}(y_t = i \mid x_t^{\mathrm{sex}} = j) = \frac{|\{y_t = i \wedge x_t^{\mathrm{sex}} = j \mid t \in 1, \ldots, N\}|}{|\{x_t^{\mathrm{sex}} = j \mid t \in 1, \ldots, N\}|}$$

What is wrong with this type of estimation?

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Probabilities as proportions of $\mathcal{D}$

$$\hat{\mathbb{P}}(y_t = i) = \frac{|\{y_t = i \mid t \in 1, \dots, N\}|}{N}$$

$$\hat{\mathbb{P}}(y_t = i \mid x_t^{\text{sex}} = j) = \frac{|\{y_t = i \wedge x_t^{\text{sex}} = j \mid t \in 1, \dots, N\}|}{|\{x_t^{\text{sex}} = j \mid t \in 1, \dots, N\}|}$$

What is wrong with this type of estimation?

- How confident should we be?

Sex and cancer

sex	cancer
F	0
M	0
F	1
F	0
F	0
F	0
M	0
F	0
M	0
F	0
M	0
M	0
F	0
M	1
M	1

We use probabilities to quantify our uncertainty.

Probabilities as proportions of $\mathcal{D}$

$$\hat{\mathbb{P}}(y_t = i) = \frac{|\{y_t = i \mid t \in 1, \ldots, N\}|}{N}$$

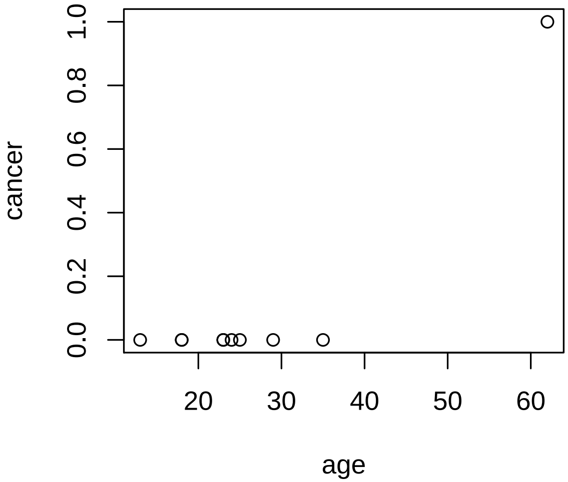
$$\hat{\mathbb{P}}(y_t = i \mid x_t^{\text{sex}} = j) = \frac{|\{y_t = i \wedge x_t^{\text{sex}} = j \mid t \in 1, \ldots, N\}|}{|\{x_t^{\text{sex}} = j \mid t \in 1, \ldots, N\}|}$$

What is wrong with this type of estimation?

- How confident should we be?
- Are there examples where it would fail??

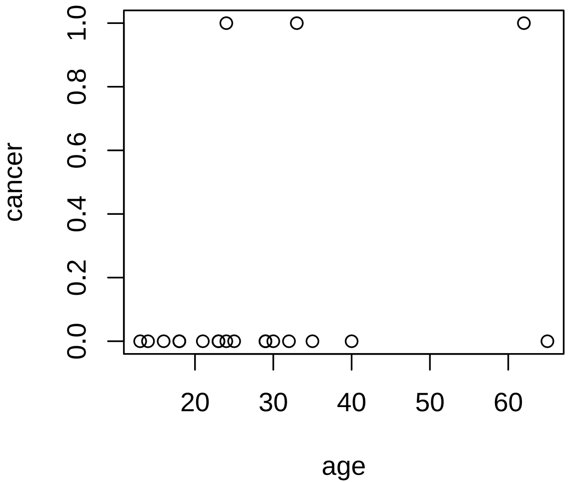
Age and cancer

age	cancer
18	0
24	0
62	1
23	0
25	0
35	0
29	0
23	0
18	0



Age and cancer

age	cancer
18	0
24	0
62	1
23	0
25	0
35	0
29	0
23	0
18	0
13	0
30	0
65	0
40	0
33	1
24	1



A small dataset

sex	smoker	cancer
F	0	0
M	0	0
F	1	1
F	0	0
F	0	0
F	0	0
M	1	0
F	0	0
M	1	0
F	0	0
M	0	0
M	0	0
F	0	0
M	1	1
M	0	1

Mixed attributes

sex	smoker	age	cancer
F	0	18	0
M	0	24	0
F	1	62	1
F	0	23	0
F	0	25	0
F	0	35	0
M	1	29	0
F	0	23	0
M	1	18	0
F	0	13	0
M	0	30	0
M	0	65	0
F	0	40	0
M	1	33	1
M	0	24	1

- Which is the "best" decision tree? Classification error vs depth/width.
- Given a criterion for what is "best", what is a good algorithm to construct it?

ID3

- An algorithm for constructing trees for binary features.
- At each step k, ID3 chooses one feature to make a decision on.
- It chooses the most "informative" feature at each step.
- It stops when no more features can be added because
  - the classification error is zero.
  - no more features are left
  - no more informative features are left

Entropy as a measure of uncertainty

Definition 2 (Entropy of a binary variable)

The entropy of a distribution with proportions p_+, p_- is

$$\mathbb{H}(p) = -p_+ \log_2 p_+ - p_- \log_2 p_- \tag{2.2}$$

Entropy as a measure of uncertainty

Definition 2 (Entropy of a binary variable)

The entropy of a distribution with proportions p_+, p_- is

$$\mathbb{H}(p) = -p_+ \log_2 p_+ - p_- \log_2 p_- \tag{2.1}$$

$$= -p_+ \log_2 p_+ - (1 - p_+) \log_2 (1 - p_+) \tag{2.2}$$

Entropy as a measure of uncertainty

Definition 2 (Entropy of a binary variable)

The entropy of a distribution with proportions p_+, p_- is

$$\begin{aligned} \mathbb{H}(p) &= -p_+ \log_2 p_+ - p_- \log_2 p_- && (2.1) \\ &= -p_+ \log_2 p_+ - (1 - p_+) \log_2 (1 - p_+) && (2.2) \end{aligned}$$

Total probability

Any probability $\mathbb{P}$ on $\{1, \ldots, K\}$ satisfies $\sum_{x=1}^{K} \mathbb{P}(x) = 1$.

Entropy as a measure of uncertainty

Definition 2 (Entropy of a binary variable)

The entropy of a distribution with proportions p_+, p_- is

$$\mathbb{H}(p) = -p_+ \log_2 p_+ - p_- \log_2 p_- \tag{2.1}$$
$$= -p_+ \log_2 p_+ - (1 - p_+) \log_2(1 - p_+) \tag{2.2}$$

Total probability

Any probability $\mathbb{P}$ on $\{1, \ldots, K\}$ satisfies $\sum_{x=1}^{K} \mathbb{P}(x) = 1$.

Definition 3 (Entropy of a discrete variable)

The entropy of a distribution $\mathbb{P}$ on alphabet $\{1, \ldots, K\}$

$$\mathbb{H}(\mathbb{P}) = -\sum_{x=1}^{K} \log \mathbb{P}(x)\, \mathbb{P}(x) = -\mathbb{E}_P \log \mathbb{P}(x).$$

Entropy as a measure of uncertainty

Definition 2 (Entropy of a discrete variable)

The entropy of a distribution $\mathbb{P}$ on alphabet $\{1, \ldots, K\}$

$$\mathbb{H}(\mathbb{P}) = -\sum_{x=1}^{K} \log \mathbb{P}(x)\, \mathbb{P}(x) = -\,\mathbb{E}_P \log \mathbb{P}(x).$$

When $\mathbb{P}$ is defined for many variables and we want to measure the entropy of some of them, it is convenient to use instead:

$$\mathbb{H}(x) = -\sum_{x=1}^{K} \log \mathbb{P}(x)\, \mathbb{P}(x) = -\,\mathbb{E}_P \log \mathbb{P}(x),$$

Conditional entropy*

What is the most informative attribute? One idea is to look at the expected reduction in entropy if we condition on that attribute.

`Example on board`

Conditional entropy*

What is the most informative attribute? One idea is to look at the expected reduction in entropy if we condition on that attribute.

Definition 3 (Conditional entropy)

The entropy of r.v. $y \in \{1, \ldots, K\}$ conditioned on r.v. $x \in \{1, \ldots, M\}$,

$$\mathbb{H}(y \mid x) = \sum_{i=1}^{M} \mathbb{P}(x = i)\mathbb{H}(y \mid x = i) \tag{2.1}$$

$$= \sum_{i=1}^{M} \mathbb{P}(x = i) \sum_{j=1}^{K} \log \mathbb{P}(y = j \mid x = i)\, \mathbb{P}(y = j \mid x = i) \tag{2.2}$$

Conditional entropy*

- The entropy of y without knowing x is $\mathbb{H}(y) = \sum_{j=1}^{K} \log[\mathbb{P}(y = j)]\,\mathbb{P}(y = k)$.
- The entropy of y when knowing $x = i$ is $\mathbb{H}(y \mid x = i) = \sum_{j=1}^{K} \log[\mathbb{P}(y = j \mid x = i)]\,\mathbb{P}(y = j)$.
- Since we don't know what value x we take, we average over the possible values: $\mathbb{H}(y \mid x) = \sum_{i=1}^{M} \log \mathbb{H}(y \mid x = i)\,\mathbb{P}(x = i)$.

Information gain*

Definition 4 (Information gain)

The information gain of variable y given x is the *expected* reduction in entropy when x becomes known.

$$\mathbb{G}(y \mid x) = \mathbb{H}(y) - \mathbb{H}(y \mid x)$$

Information gain*

Definition 4 (Information gain)

The information gain of variable y given x is the *expected* reduction in entropy when x becomes known.

$$\mathbb{G}(y \mid x) = \mathbb{H}(y) - \mathbb{H}(y \mid x)$$

For ID3, we use the empirical distribution of x, y from $\mathcal{D}$ (the observed proportions) as $\mathbb{P}$.

Shorthand notation for classification

Since we're only interested in the entropy of labels y, we write

- $\mathbb{H}(\mathcal{D})$ for the entropy of y wrt the empirical distribution.
- $\mathbb{H}(\mathcal{D}_{a=v})$ when attribute a takes the value v.
- $\mathbb{G}(\mathcal{D}, a)$ for the information gain conditioned on a.

ID3($\mathcal{D}, \mathcal{A}$)

Make new node.

if $\exists i : y = i \forall (x, y) \in \mathcal{D}$, or $\mathcal{A} = \emptyset$ `// nothing to do` **then**

  Set label to $\arg\max_i |\{y = i, (x, y) \in \mathcal{D} \mid |\}$ `// use maximum class`

end if

$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{G}(\mathcal{D}, a)$.

for $v \in \mathcal{V}_{a^*}$ **do**

  Make a new branch v

  if $\mathcal{D}_{a^*=v} \neq \emptyset$ **then**

    ID3($\mathcal{D}_{a^*=v}, \mathcal{A} - \{a^*\}$)

  end if

end for

ID3($\mathcal{D}, \mathcal{A}$)

Make new node.

if $\exists i : y = i \forall (x, y) \in \mathcal{D}$, or $\mathcal{A} = \emptyset$ `// nothing to do` **then**

  Set label to $\arg\max_i |\{y = i, (x, y) \in \mathcal{D} \mid |\}$ `// use maximum class`

end if

$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{G}(\mathcal{D}, a)$.

for $v \in \mathcal{V}_{a^*}$ **do**

  Make a new branch v

  if $\mathcal{D}_{a^*=v} \neq \emptyset$ **then**

    ID3($\mathcal{D}_{a^*=v}, \mathcal{A} - \{a^*\}$)

  end if

end for

ID3($\mathcal{D}, \mathcal{A}$)

Make new node.
if $\exists i : y = i \forall (x, y) \in \mathcal{D}$, or $\mathcal{A} = \emptyset$ `// nothing to do` **then**
    Set label to $\arg\max_i |\{y = i, (x, y) \in \mathcal{D} \mid |\}$ `// use maximum class`
end if
$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{G}(\mathcal{D}, a)$.
for $v \in \mathcal{V}_{a^*}$ **do**
    Make a new branch v
    if $\mathcal{D}_{a^*=v} \neq \emptyset$ **then**
        ID3($\mathcal{D}_{a^*=v}, \mathcal{A} - \{a^*\}$)
    end if
end for

ID3($\mathcal{D}, \mathcal{A}$)

Make new node.
if $\exists i : y = i \forall (x,y) \in \mathcal{D}$, or $\mathcal{A} = \emptyset$ `// nothing to do` **then**
  Set label to $\arg\max_i |\{y = i, (x,y) \in \mathcal{D} \mid |\}$ `// use maximum class`
end if
$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{G}(\mathcal{D}, a)$.
for $v \in \mathcal{V}_{a^*}$ **do**
  Make a new branch v
  if $\mathcal{D}_{a^*=v} \neq \emptyset$ **then**
    ID3($\mathcal{D}_{a^*=v}, \mathcal{A} - \{a^*\}$)
  end if
end for

$\mathtt{ID3}(\mathcal{D}, \mathcal{A})$

Make new node.
if $\exists i : y = i \forall (x, y) \in \mathcal{D}$, or $\mathcal{A} = \emptyset$ `// nothing to do` **then**
  Set label to $\arg\max_i |\{y = i, (x, y) \in \mathcal{D} \mid |\}$ `// use maximum class`
end if
$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{G}(\mathcal{D}, a)$.
for $v \in \mathcal{V}_{a^*}$ **do**
  Make a new branch v
  if $\mathcal{D}_{a^*=v} \neq \emptyset$ **then**
    $\mathtt{ID3}(\mathcal{D}_{a^*=v}, \mathcal{A} - \{a^*\})$
  end if
end for

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

D

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$\begin{tikzcd} & D_S \\ D \arrow[ur] \arrow[dr] & \\ & D_{\neg S} \end{tikzcd}$$

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[dr] & & \\
 & D_{\neg S} &
\end{tikzcd}$$

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$
\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_{\neg S} \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S}
\end{tikzcd}
$$

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_{\neg S} \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S}
\end{tikzcd}$$

D

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$
\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_{\neg S} \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S}
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
 & D_F \\
D \arrow[ur] \arrow[dr] & \\
 & D_M
\end{tikzcd}
$$

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$
\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_{\neg S} \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S}
\end{tikzcd}
\qquad
\begin{tikzcd}
 & & D_{F,S} \\
 & D_F \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & \\
 & D_M &
\end{tikzcd}
$$

ID3 example

Smoking	Sex	Cancer
Yes	Male	Yes
No	Male	No
Yes	Female	No
No	Female	No

$$
\begin{tikzcd}
 & & D_{F,S} \\
 & D_S \arrow[ur] \arrow[dr] & \\
 & & D_{M,S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_{\neg S} \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S}
\end{tikzcd}
\qquad
\begin{tikzcd}
 & & D_{F,S} \\
 & D_F \arrow[ur] \arrow[dr] & \\
 & & D_{F,\neg S} \\
D \arrow[uur] \arrow[ddr] & & \\
 & & D_{M,\neg S} \\
 & D_M \arrow[ur] \arrow[dr] & \\
 & & D_{M,S}
\end{tikzcd}
$$

Questions about ID3

- After ID3 ends, are all training examples classified correctly?

Questions about ID3

- After ID3 ends, are all training examples classified correctly?
- Does the order in which we add features matter for the training classification error?

Questions about ID3

- After ID3 ends, are all training examples classified correctly?
- Does the order in which we add features matter for the training classification error?
- Does the order matter for the testing classification error?

Questions about ID3

- After ID3 ends, are all training examples classified correctly?
- Does the order in which we add features matter for the training classification error?
- Does the order matter for the testing classification error?
- Does the order matter if we make the tree shorter?

Questions about ID3

- After ID3 ends, are all training examples classified correctly?
- Does the order in which we add features matter for the training classification error?
- Does the order matter for the testing classification error?
- Does the order matter if we make the tree shorter?
- If examples are inconsistent, how can we achieve perfect classification? *Hint: use data augmentation*

Generalising decision trees

- We can think of more general versions of ID3.
- Can work with non-binary features.
- Use other criteria to split (e.g. the expected reduction in classification error)
- Can also do regression.

The C4.5 algorithm

Identical to ID3 apart from dealing with numeric variables.

Numeric attribute splitting

- For each attribute a
- Look at all possible splitting points x
- Calculate $\mathbb{G}$ for each combination a, x.
- Use that!

Expected reduction in classification error*

The classification error for a given subset $\mathcal{D}_i$ of the data $\mathcal{D}$ is the proportion of labels not equal to the most frequent label:

Example 5

Name	Smoking
Silvie	Yes
Yiannis	Yes
Marie	No
Claudia	Yes
Jonas	Yes
Andrei	No
Keisuke	Yes
Yamada	Yes
Lee	No

Expected reduction in classification error*

The classification error for a given subset $\mathcal{D}_i$ of the data $\mathcal{D}$ is the proportion of labels not equal to the most frequent label:

Example 5

Name	Smoking
Silvie	Yes
Yiannis	Yes
Marie	No
Claudia	Yes
Jonas	Yes
Andrei	No
Keisuke	Yes
Yamada	Yes
Lee	No

Question:What is the classification error of the best fixed decision for the highlighted subset?

Expected reduction in classification error*

The classification error for a given subset $\mathcal{D}_i$ of the data $\mathcal{D}$ is the proportion of labels not equal to the most frequent label: Question: What is the classification error for the best fixed decision for the highlighted subset over the remaining dataset?

Expected reduction in classification error*

The classification error for a given subset $\mathcal{D}_i$ of the data $\mathcal{D}$ is the proportion of labels not equal to the most frequent label:

Definition 5 (Classification error (of a fixed decision rule) for a set $\mathcal{D}_i$)

$$\epsilon(\mathcal{D}_i) \triangleq \frac{|\{y \neq y^*(\mathcal{D}_i) \mid (x, y) \in \mathcal{D}_i|\}}{|\mathcal{D}_i|}, \tag{3.1}$$

$$y^*(\mathcal{D}_i) \triangleq \arg\max_{k \in Y} |\{y = k \mid (x, y) \in \mathcal{D}_i\}| \tag{3.2}$$