582631 Introduction to Machine Learning, Fall 2016
Exercise set V
Example solutions

1.

(a) Consider a first split into on the last week's decision tree (Figure 1). A set D is the whole data set, D_1 is an upper half of the goal R_1, and $D_2 = \bigcup_{i=2}^{6} R_i$ is the lower half of the goal. Denote goals by 1, and saved shots by 0.

Now the proportions of the goals on the aforementioned sets are

$$\begin{aligned} p_1(D) &= \frac{k}{|D|} = \frac{130}{168}, \\ p_1(D_1) &= \frac{k_1}{|D_1|} = \frac{49}{52}, \\ p_1(D_2) &= \frac{k_2}{|D_2|} = \frac{81}{116}. \end{aligned}$$

Gini indices for these sets are

$$\begin{aligned} Q(D) &= p_1(D)(1 - p_1(D)) + p_0(D)(1 - p_0(D)) \\ &= 2p_1(D)(1 - p_1(D)) = 2 \cdot \frac{130}{168} \cdot \frac{38}{168} \approx 0.350 \\ Q(D_1) &= 2p_1(D_1)(1 - p_1(D_1)) = 2 \cdot \frac{49}{52} \cdot \frac{3}{52} \approx 0.109 \\ Q(D_2) &= 2p_1(D_2)(1 - p_1(D_2)) = 2 \cdot \frac{81}{116} \cdot \frac{35}{116} \approx 0.421 \end{aligned}$$

Now the gain using the Gini index as an impurity measure can be computed as

$$\begin{aligned} \text{gain}(\{D_1, D_2\}) &= Q(D) - \left(\frac{|D_1|}{|D|} Q(D_1) + \frac{|D_2|}{|D|} Q(D_2) \right) \\ &= 0.350 - \left(\frac{52}{168} \cdot 0.109 + \frac{116}{168} \cdot 0.421 \right) \approx 0.025. \end{aligned}$$

Values of the gain (using Gini) for the rest of the splits are shown in the following table:

Split	k_1	k_2	$\lvert D_1 \rvert$	$\lvert D_2 \rvert$	$Q(D_1)$	$Q(D_2)$	$Q(D)$	$\text{gain}(D_1, D_2)$
R_1	49	81	52	116	0.350	0.109	0.421	0.025
R_2	21	60	21	95	0.421	0.000	0.465	0.040
R_3	21	39	27	68	0.465	0.346	0.489	0.017
R_4	16	23	26	42	0.489	0.473	0.496	0.002
R_5	12	11	15	27	0.496	0.320	0.483	0.071

Values of the cross-entropy (all logarithms are 2-based) for the sets in the first split are

$$\begin{aligned}
Q(D) &= -p_1(D)\log p_1(D) - p_0(D)\log p_0(D) \\
&= -\frac{130}{168}\log\frac{130}{168} - \frac{38}{168}\log\frac{38}{168} \approx 0.771, \\
Q(D_1) &= -p_1(D_1)\log p_1(D_1) - p_0(D_1)\log p_0(D_1), \\
&= -\frac{49}{52}\log\frac{49}{52} - \frac{3}{52}\log\frac{3}{52} \approx 0.318, \\
Q(D_2) &= -p_1(D_2)\log p_1(D_2) - p_0(D_2)\log p_0(D_2), \\
&= -\frac{81}{116}\log\frac{81}{116} - \frac{36}{116}\log\frac{36}{116} \approx 0.883.
\end{aligned}$$

Now the gain using the cross entropy as an impurity measure can be computed as

$$\begin{aligned}
\text{gain}(\{D_1, D_2\}) &= Q(D) - \left(\frac{|D_1|}{|D|}Q(D_1) + \frac{|D_2|}{|D|}Q(D_2)\right) \\
&= 0.771 - \left(\frac{52}{168}\cdot 0.318 + \frac{116}{168}\cdot 0.883\right) \approx 0.063.
\end{aligned}$$

Values of the gain (using cross entropy) for the rest of the splits are shown in the following table:

Split	k_1	k_2	$\lvert D_1\rvert$	$\lvert D_2\rvert$	$Q(D_1)$	$Q(D_2)$	$Q(D)$	$\text{gain}(D_1, D_2)$
R_1	49	81	52	116	0.771	0.318	0.883	0.063
R_2	21	60	21	95	0.883	0.000	0.949	0.106
R_3	21	39	27	68	0.949	0.764	0.984	0.028
R_4	16	23	26	42	0.984	0.961	0.993	0.003
R_5	12	11	15	27	0.993	0.722	0.975	0.109

(b) Denote by k the number of observations of class $Y = 1$ in the set D, and by k_1 and k_2 in the subsets D_1 and D_2, respectively; $k = k_1 + k_2$ because $\{D_1, D_2\}$ is a partition of D. Because by assumption $Y = 0$ is a majority class in D, D_1 and D_2 :

$$\begin{aligned}
Q(D) &= 1 - \max_c p_c(D) = p_1(D) = \frac{k}{|D|}, \\
Q(D_1) &= 1 - \max_c p_c(D_1) = p_1(D_1) = \frac{k_1}{|D_1|} \text{ and} \\
Q(D_2) &= 1 - \max_c p_c(D_2) = p_1(D_2) = \frac{k_2}{|D_2|.}
\end{aligned}$$

Substituting these into the formula for gain, we see that the gain is zero:

$$\begin{aligned}
\text{gain}(\{D_1, D_2\}) &= Q(D) - Q(\{D_1, D_2\}) \\
&= Q(D) - \left(\frac{|D_1|}{|D|}Q(D_1) + \frac{|D_2|}{|D|}Q(D_2)\right) \\
&= \frac{k}{|D|} - \left(\frac{|D_1|}{|D|}\frac{k_1}{|D_1|} + \frac{|D_2|}{|D|}\frac{k_2}{|D_2|}\right) \\
&= \frac{k}{|D|} - \frac{k_1 + k_2}{|D|} = 0.
\end{aligned}$$

The assumption that the majority class is the same in both D_1 and D_2 is essential for the result: otherwise the gain will be positive.

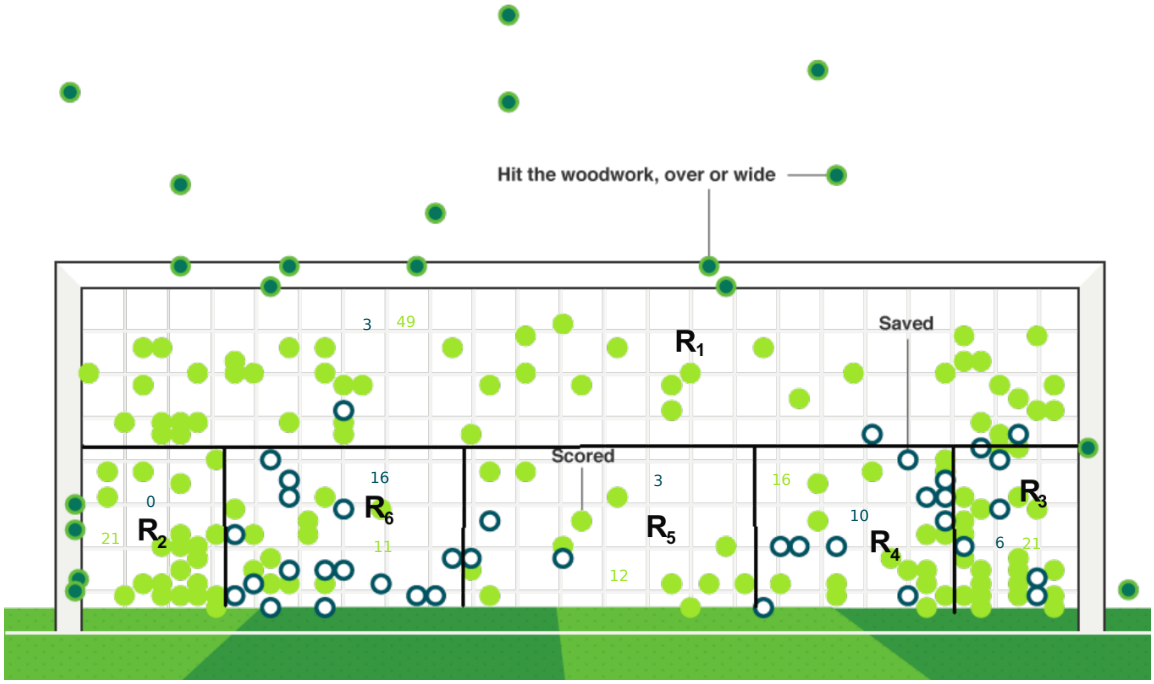

Figure 1: Penalties shot in World cup up to South Africa 2010

(c) When written as a function of $p_0(D)$, the proportion of class $Y = 0$ in the set D, both Gini and cross-entropy are concave functions. This means that a straight line segment connecting any two points on the curve is always below the curve.

Observe first that $p(D)$ can be written as a convex combination of $p_0(D_1)$ and $p_0(D_2)$:

$$\begin{aligned} p_0(D) &= \frac{k}{|D|} = \frac{k_1 + k_2}{|D|} \\ &= \frac{|D_1|}{|D|} \cdot \frac{k_1}{|D_1|} + \frac{|D_2|}{|D|} \cdot \frac{k_2}{|D_2|} \\ &= \alpha p_0(D_1) + (1 - \alpha) p_0(D_2), \end{aligned} \tag{1}$$

where

$$0 \leq \alpha = \frac{|D_1|}{|D|} \leq 1.$$

Now consider the definition of *gain*:

$$\text{gain}(D_1, D_2) = Q(D) - \left(\frac{|D_1|}{|D|} Q(D_1) + \frac{|D_2|}{|D|} Q(D_2) \right) = Q(D) - (\alpha Q(D_1) + (1 - \alpha) Q(D_2)) \,.$$

The first term, $Q(D)$ can be written using Eq. (1):

$$Q(D) \equiv Q(p_0(D)) = Q(\alpha p_0(D_1) + (1 - \alpha) p_0(D_2)).$$

Thus the gain is positive if and only if

$$Q(D) = Q(\alpha p_0(D_1) + (1 - \alpha) p_0(D_2)) > \alpha Q(D_1) + (1 - \alpha) Q(D_2).$$

This is exactly what Jensen's inequality implies for concave functions.[1]

[1] Often, Jensen's inequality is given in a form that states the opposite inequality for *convex* functions, but this the same thing since for a concave function, f, we obtain a convex function $-f$, and the direction of the inequality can be flipped back by multiplying both sides by -1.

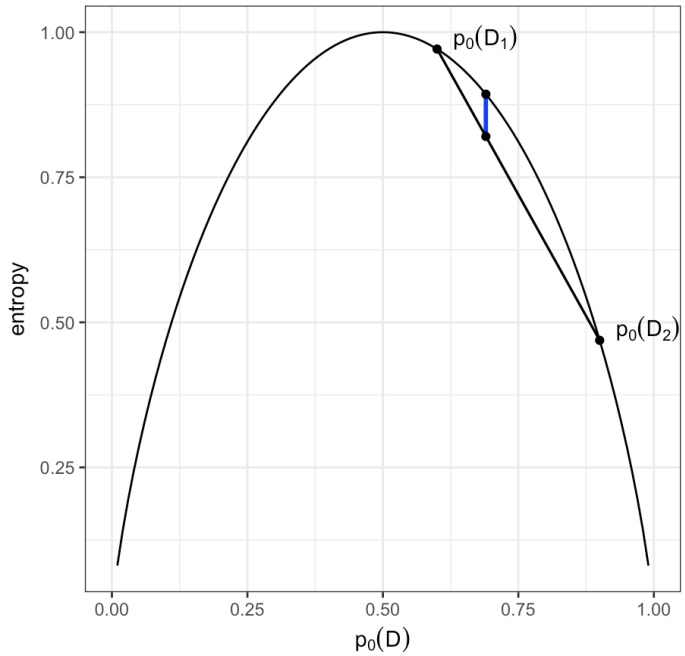

Figure 2: Entropy and Jensen's inequality with points $(p_0(D_1), Q(D_1))$ and $(p_0(D_2), Q(D_2))$ connected by a straight line. At the point where $p_0(D) = \alpha p_0(D_1) + (1-\alpha)p_0(D_2)$, the line segment connecting the end points of the line segment is below the entropy curve. The difference shown as a blue line gives the gain of the split.

By solving where this inequality holds as an equality, we see that the only solution, and so also the only point where gain is zero is $p_0(D_1) = p_0(D_2)$ for both Gini and cross-entropy – or in fact, if $\alpha \in \{0, 1\}$, in which case either $Q(D_1)$ or $Q(D_2)$ is undefined since one side of the split would be empty.

2.

(c) The joint probability for the class value $Y = 0$ and the feature vector value $\mathbf{X} = (0, 0)$ can be computed as

$$P(Y = 0, X_1 = 0, X_2 = 0) = P(Y = 0)P(X_1 = 0, X_2 = 0|Y = 0) = 0.4 \cdot 0.2 = 0.08.$$

The joint probabilities for the rest of the values are computed with a same formula, and can be found on the following table:

	$Y = 0$	$Y = 1$	$Y = 2$
$P(Y = y, X_1 = 0, X_2 = 0)$	0.08	0.18	0.03
$P(Y = y, X_1 = 1, X_2 = 0)$	0.04	0.03	0.12
$P(Y = y, X_1 = 0, X_2 = 1)$	0.16	0.03	0.09
$P(Y = y, X_1 = 1, X_2 = 1)$	0.08	0.03	0.00
$P(Y = y, X_1 = 0, X_2 = 2)$	0.00	0.03	0.06
$P(Y = y, X_1 = 1, X_2 = 2)$	0.04	0.00	0.00

The value of the joint probability for the most likely class (and so the class predicted by the Bayes optimal classifier) $h(x_1, x_2)$ for each feature vector value $\mathbf{X} = (x_1, x_2)$ is highlighted.

Since the Bayes error rate is the probability that a true class value is not the most likely class value, it can be computed by summing the probability of this event over the feature vector values $\mathbf{X} = (x_1, x_2)$:

$$\begin{aligned} P(h(X_1, X_2) \neq Y) &= 1 - P(h(X_1, X_2) = Y) \\ &= 1 - \sum_{x_1=0}^{1} \sum_{x_2=0}^{2} P(Y = h(x_1, x_2), X_1 = x_1, X_2 = x_2) \\ &= 1 - (0.18 + 0.12 + 0.16 + 0.08 + 0.06 + 0.04) \\ &= 0.36. \end{aligned}$$