| Formula | Variables | Example |
|---|---|---|
| $E(D) = \sum_{i=1}^{i=c} -p_i \log_2(p_i)$ | $D$: dataset with $c$ classes<br>$p_i$: probability of class $i$ | A set of training data contains 9 yes and 11 no.<br>$E(D) = -\frac{9}{20}\log_2\left(\frac{9}{20}\right) - \frac{11}{20}\log_2\left(\frac{11}{10}\right) = 0.9928$ |
| $Gain(D, A) = E(D) - \sum_{j=1}^{j=v} \frac{\lvert D_j \rvert}{\lvert D \rvert} E(D_j)$ | $D$: dataset with $c$ classes<br>$A$: feature<br>$\lvert D \rvert$: number of elements in $D$<br>$\lvert D_j \rvert$: $\{x \lvert x \in D \wedge A\ has\ value\ j\}$ | Now consider feature $T$ with values $A, G, P$,<br>$D_A = 6, D_G = 9, D_P = 5$.<br>$Gain(D, T) = E(D) - \frac{6}{20} \cdot E(D_A) - \cdots$ |
| $w_o + w_1 x_1 + w_2 x_2 = 0$ | This is the general form of a classification line.<br>If $w_0 + w_1 x_1 + w_2 x_2 > 0$, the class is 1, otherwise as 0. | |
| $(w_o^{new}, w_1^{new}, w_2^{new}) = (w_0^{old}, w_1^{old}, w_2^{old}) + \alpha(1, a_1, a_2)$ | $(a_1, a_2)$: point that is 1 but classified as 0. | You're not a retard. |
| $(w_o^{new}, w_1^{new}, w_2^{new}) = (w_0^{old}, w_1^{old}, w_2^{old}) - \alpha(1, b_1, b_2)$ | $(b_1, b_2)$: point that is 0 but classified as 1. | You're not a retard. |
| $accuracy = \frac{a+d}{a+b+c+d}$<br>$error\ rate = \frac{b+c}{a+b+c+d}$<br>$recall\ C_1 = \frac{a}{a+b}$<br>$precision\ C_1 = \frac{a}{a+c}$<br>$recall\ C_2 = \frac{d}{c+d}$<br>$precision\ C_2 = \frac{d}{b+d}$ | <table><tr><td></td><td></td><td>Predicated class</td><td></td></tr><tr><td></td><td></td><td>$C_1$</td><td>$C_2$</td></tr><tr><td>Actual class</td><td>$C_1$</td><td>$a$</td><td>$b$</td></tr><tr><td></td><td>$C_2$</td><td>$c$</td><td>$d$</td></tr></table><br>High precision means little false positives, while high recall means little false negatives. | |
| $P(A\|B) = \frac{P(B\|A) \cdot P(A)}{P(B)}$ | $P(A\|B)$: probability of $A$ given $B$<br>$P(A)$: probability of $A$ (…) | |
| $P(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$ | | No idea at all.<br>Bernouilli trial it's called. |

**Exercises 3.1.2 Information Gain**

1. If you know only feature A, you still know nothing: for every given value for feature A, the probability of a green or red ball is still 0.5.
   However, if you know only feature B, suddenly the probabilities shift a bit. For example, if B = 1, then the ball is green, and if B = 4, then the ball is red, et cetera. The odds of guessing the color right when knowing feature B are higher than when you know feature A. So I'd choose feature B.
2. I pay 1 and get either 1.5 if I win, and 0 if I lose.
   When I know $B$, the probability that I correctly guess the color increases.
   To be precise, this is the average of the probabilities for each value of $B$.
   So, $p(win|I\ know\ B) = \frac{1+\frac{4}{6}+\frac{4}{6}+1+\frac{5}{6}+\frac{5}{6}}{6} \cong .833$.

1. The set of examples for which $T$ has a value $P$, called $D_P$, consists of 9 examples, of which 2 have a positive assessment. Hence,
   $E(D_P) = -\frac{2}{9}\log_2\left(\frac{2}{9}\right) - \frac{7}{9}\log_2\left(\frac{7}{9}\right) = 0.7642$
2. The set of examples for which $T$ has a value $G$, called $D_G$, consists of 5 examples, that all have a positive assessment. By definition this means that $E(D_G) = 0$.

**Exercises 3.2.1 Some Mathematics**

Given the classification $3 + 2x_1 - 2x_2 = 0$, the feature point (2,3) actually means $x_1 = 2$ and $x_2 = 3$. Replacing that gives $3 + 2 \cdot 2 - 2 \cdot 3 = 3 + 4 - 6 = 1 > 0$, so the feature point will be classified as 1.

**Exercises 3.2.2 The Intuition behind Gradient Descent**

We again consider $3 + 2x_1 - 2x_2 = 0$, or $w = (3,2,-2)$, and the feature (2,3) having a class 0. We also say that the learning rate $\alpha = 0.5$.

The feature is falsely seen as class 1, so we have to adapt $w$ like the following:

$$(w^0_{new}, w^1_{new}, w^2_{new}) = (w^0_{old}, w^1_{old}, w^2_{old}) - \alpha(1,2,3)$$
$$= (3,2,-2) - 0.5(1,2,3)$$
$$= (3,2,-2) - (.5,1,1.5)$$
$$= (2.5,1,-3.5)$$

Now the feature point $(2,3)$ will be classified like so:
$2.5 + 1 \cdot 3 - 3.5 \cdot 2 = 2.5 + 3 - 7 = -1.5 < 0$, so the feature is classified as 0. The error is corrected.

**Exercises 4.2 Confusion matrix**

1. In this confusion matrix, $a = 80, b = 10, c = 20, d = 50$.
   a) $accuracy = \frac{a+d}{a+b+c+d} = \frac{80+50}{80+10+20+50} = \frac{130}{160}$.
   b) $precision\ for\ C_1 = \frac{a}{a+c} = \frac{80}{80+20} = \frac{80}{100}$.
   c) $recall\ for\ C_2 = \frac{b}{b+a} = \frac{10}{10+80} = \frac{10}{90}$.

$$\frac{50}{70}$$

2. High precision means little false positives, while high recall means little false negatives. So something that needs as little as possible false positives needs high precision, whilst something that needs as little as possible false negatives needs high recall.

When building table:

Columns: $H$, $P(H)$, $P(D|H)$, $P(D|H) \cdot P(H)$, $p(H|D)$
$P(H|D)$=$\frac{P(D|H)\cdot P(H)}{\sum P(D|H)\cdot P(H)}$