Connections to machine-learning

Patrick Rubin-Delanchy
patrick.rubin-delanchy@bristol.ac.uk

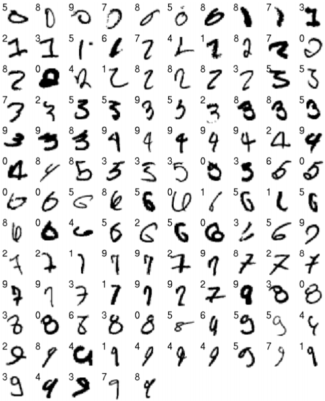
December 17, 2018

The classification setup

A binary classifier on $\mathbb{R}^d$ is a (Borel measurable) function $C : \mathbb{R}^d \to \{0, 1\}$. (Cannings and Samworth, 2017)

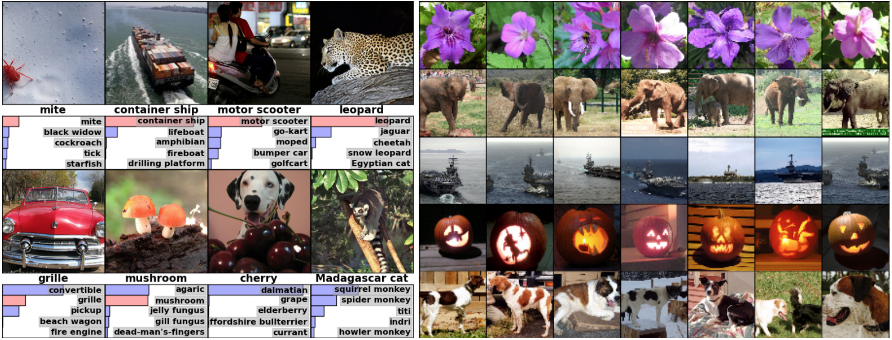
In words, given a vector in $\mathbb{R}^d$ describing a particular object (e.g. a picture), called a *feature vector*, a classifier predicts whether the object belongs to class 0 (e.g. cat) or class 1 (e.g. dog).

A famous example ($C : \mathbb{R}^d \to \{0, \ldots, 9\}$)



Source: Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation* 18.7 (2006): 1527-1554.

A famous example: $(C : \mathbb{R}^d \to \{0, \ldots, ?\})$



Source: Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “Imagenet classification with deep convolutional neural networks.” *Advances in neural information processing systems*. 2012.

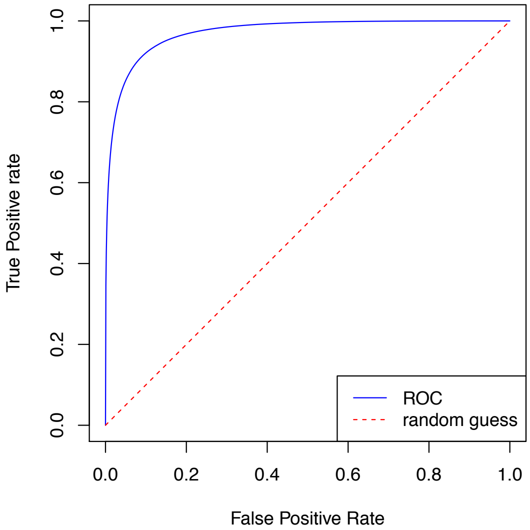
In practice, the typical output of a binary classifier (e.g. Random Forests, neural networks, support vector machines) is a single number indicating a degree of preference for class 1 (higher favouring class 1).

This allows the user to choose a threshold, τ, that achieves the desired balance of false positives versus false negatives.

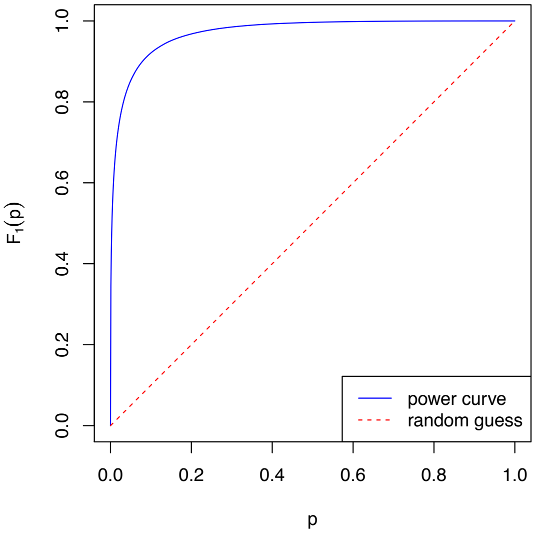
In particular there are many practical examples where the cost of a false positive is deemed high relative to a false negative, e.g. in criminal trials.

Furthermore, there are often many more instances of one class over another. In cyber-security, it is common to talk about being "swamped with false positives".

Receiver Operating Characteristic curve



Receiver Operating Characteristic curve



Area under the curve

A common (but not necessarily unproblematic) measure of performance of a classifier is the "area under the curve" (AUC) of the ROC.

This is a number which in principle lies within $[0, 1]$, but more typically lies within $[1/2, 1]$ (since an AUC $< 1/2$ suggests the classifier is worse than random).

The AUC is commonly interpreted as the probability of a random instance of class 1 being ranked higher than a random instance of class 0.

By the above discussion, define a classifier, instead, to be a function $C : \mathbb{R}^d \to \mathbb{R}$, and reframe the classification setup as a hypothesis test:

$$H_0 : X \sim F_0, \\ \text{versus,} \\ H_1 : X \sim F_1,$$

where $X \in \mathbb{R}^d$, using the function C as a test, i.e., $T = C(X)$.

We have already seen that the ROC is the power curve of this test, or the cumulative distribution function F_1 of the p-value under the alternative. Now, let $\mathbb{E}_1$ denote the expectation of the p-value under the alternative.

By the above discussion, define a classifier, instead, to be a function $C : \mathbb{R}^d \to \mathbb{R}$, and reframe the classification setup as a hypothesis test:

$$H_0 : X \sim F_0,$$

versus,

$$H_1 : X \sim F_1,$$

where $X \in \mathbb{R}^d$, using the function C as a test, i.e., $T = C(X)$.

We have already seen that the ROC is the power curve of this test, or the cumulative distribution function F_1 of the p-value under the alternative. Now, let $\mathbb{E}_1$ denote the expectation of the p-value under the alternative.

Theorem (Expected p-value)

$$AUC = 1 - \mathbb{E}_1(P)$$

A digression

Let X be a non-negative random variable and define $S(t) = \mathbb{P}(X \geq t)$. Then,

$$\mathbb{E}(X) = \int_0^\infty S(t)\mathrm{d}t.$$

Proof.

We have

$$\int_0^\infty S(t)\mathrm{d}t$$

□

A digression

Let X be a non-negative random variable and define $S(t) = \mathbb{P}(X \geq t)$. Then,

$$\mathbb{E}(X) = \int_0^\infty S(t)\mathrm{d}t.$$

Proof.

We have

$$\int_0^\infty S(t)\mathrm{d}t = \int_0^\infty \mathbb{E}\{\mathbb{I}(X \geq t)\}\mathrm{d}t$$

□

A digression

Let X be a non-negative random variable and define $S(t) = \mathbb{P}(X \geq t)$. Then,

$$\mathbb{E}(X) = \int_0^\infty S(t)\mathrm{d}t.$$

Proof.

We have

$$\begin{aligned}\int_0^\infty S(t)\mathrm{d}t &= \int_0^\infty \mathbb{E}\{\mathbb{I}(X \geq t)\}\mathrm{d}t \\ &= \mathbb{E}\left\{\int_0^\infty \mathbb{I}(X \geq t)\mathrm{d}t\right\}\end{aligned}$$

□

A digression

Let X be a non-negative random variable and define $S(t) = \mathbb{P}(X \geq t)$. Then,

$$\mathbb{E}(X) = \int_0^\infty S(t)\mathrm{d}t.$$

Proof.

We have

E()=1*S+0*(1-S)

$$\begin{aligned}\int_0^\infty S(t)\mathrm{d}t &= \int_0^\infty \mathbb{E}\{\mathbb{I}(X \geq t)\}\mathrm{d}t \\ &= \mathbb{E}\left\{\int_0^\infty \mathbb{I}(X \geq t)\mathrm{d}t\right\} \\ &= \mathbb{E}(X).\end{aligned}$$

↲ ?

□

The theorem easily follows from

$$\mathsf{AUC} = \int_0^1 F_1(p)\mathrm{d}p = 1 - \int_0^1 S_1(p) = 1 - \mathbb{E}_1(P).$$

The theorem easily follows from

$$\mathsf{AUC} = \int_0^1 F_1(p)\mathrm{d}p = 1 - \int_0^1 S_1(p) = 1 - \mathbb{E}_1(P).$$

Now let $P_0 \sim \mathrm{uniform}[0,1]$, with density $f_0 = 1$ and $P_1 \sim F_1$, with density f_1. Then, if P_0 and P_1 are independent:

$$\mathbb{P}(P_1 \leq P_0)$$

The theorem easily follows from

$$\mathrm{AUC} = \int_0^1 F_1(p)\mathrm{d}p = 1 - \int_0^1 S_1(p) = 1 - \mathbb{E}_1(P).$$

Now let $P_0 \sim \text{uniform}[0,1]$, with density $f_0 = 1$ and $P_1 \sim F_1$, with density f_1. Then, if P_0 and P_1 are independent:

$$\mathbb{P}(P_1 \leq P_0) = \int_0^1 \mathbb{P}(P_1 \leq P_0 \mid P_0 = p) f_0(p)\mathrm{d}p$$

The theorem easily follows from

$$\mathsf{AUC} = \int_0^1 F_1(p)\mathrm{d}p = 1 - \int_0^1 S_1(p) = 1 - \mathbb{E}_1(P).$$

Now let $P_0 \sim \mathrm{uniform}[0,1]$, with density $f_0 = 1$ and $P_1 \sim F_1$, with density f_1. Then, if P_0 and P_1 are independent:

$$\begin{aligned}\mathbb{P}(P_1 \leq P_0) &= \int_0^1 \mathbb{P}(P_1 \leq P_0 \mid P_0 = p) f_0(p)\mathrm{d}p \\ &= \int_0^1 F_1(p)\mathrm{d}p\end{aligned}$$

The theorem easily follows from

$$\mathrm{AUC} = \int_0^1 F_1(p)\mathrm{d}p = 1 - \int_0^1 S_1(p) = 1 - \mathbb{E}_1(P).$$

Now let $P_0 \sim \text{uniform}[0, 1]$, with density $f_0 = 1$ and $P_1 \sim F_1$, with density f_1. Then, if P_0 and P_1 are independent:

$$\begin{aligned}\mathbb{P}(P_1 \leq P_0) &= \int_0^1 \mathbb{P}(P_1 \leq P_0 \mid P_0 = p) f_0(p)\mathrm{d}p \\ &= \int_0^1 F_1(p)\mathrm{d}p \\ &= 1 - \mathbb{E}(P_1),\end{aligned}$$

explaining the afore-mentioned interpretation of the AUC.

Precision, recall, etc

Several other measures of classification performance are popular, and are linked with earlier notions we have learned.

1. Precision: fraction of true positives among positives

Precision, recall, etc

Several other measures of classification performance are popular, and are linked with earlier notions we have learned.

1. Precision: fraction of true positives among positives
   $= 1 -$ empirical false discovery rate

Precision, recall, etc

Several other measures of classification performance are popular, and are linked with earlier notions we have learned.

1. Precision: fraction of true positives among positives
   $= 1 -$ empirical false discovery rate
2. Recall: fraction of true positives among all true instances

An issue with these measures is that they are not invariant to the relative numbers of instances of each class, and can therefore make comparing classifiers confusing.