CISC 271 Class 25

Odds of Occurrence and Probability

Text Correspondence: $\sim$

Main Concepts:

- *Odds: ratio of likelihood of an event*
- *Probability: inverse function of odds*
- *Logistic function: map from score to probability*
- *Logistic regression: use $p(d)$ instead of d*

Sample Problem, Machine Inference: How can we compute the probability of classification for a data vector?

We now know at least one way to perform binary classification: using linear separation. We might process a data matrix X to find a separating hyperplane $\mathbb{H}$; this might be unsupervised clustering or might be supervised and use labels for the observations. Consider a new data observation $\underline{x}$; we can classify this observation as a member of Class +1 based on a hyperplane $\mathbb{H}$ that is specified by a non-zero normal vector $\vec{m}$ and a bias b. We can do this by finding the signed distance d from $\vec{a}$ to the hyperplane $\mathbb{H}$ as a score

$$d(\underline{x}; \vec{n}, c) \stackrel{\text{def}}{=} \underline{x}\vec{n} + c \tag{25.1}$$

We assign an observation $\underline{x}$ to Class +1 if and only if $d(\underline{x}; \vec{n}, c) \geq 0$. Now, we will try answer a more general question: what is the *probability* that an observation $\underline{x}$ has the label+1?

For us, a probability is a measure of the likelihood of an outcome. The outcome is whether an observation $\underline{x}$ has the label +1. A probability is a real number between 0 and 1; because a probability is usually written as p, and because we will not model perfect outcomes, we require

$$p \in (0, 1)$$

We will not explore interpretations of probability, or probability density functions, which are extensive fields of study. Interested students might investigate offerings in mathematics and statistics to learn more about these important topics.

We begin by considering a common statement made in gambling and other practical applications of probability theory. When someone states "the odds are 50:50", we understand this as

© R E Ellis 2025

meaning that the probability of some outcome is 50%; we infer that the probability of the outcome being something else is also 50%, so the ratio of the probabilities is

$$50/50 = 0.5/0.5$$

In binary classification, we are understanding the statement to mean that the likelihood that an observation has the label +1 is the same as the likelihood that an observation has the label -1.

When someone states "the odds are 9 to 1", we understand this in binary classification as meaning that the probability that an observation has the label +1 is 9 times as likely as the probability that an observation has the label -1. The two probabilities must add to 1, because in binary classification no other outcome is possible. If we write that the probability of an outcome is p, then we mean that the probability of an observation $\underline{x}$ has the label +1 is p. Because the probability that an observation $\underline{x}$ *either* has the label +1 or the label -1 must be 100%, or 1, we can infer that the probability that $\underline{x}$ has the label -1 is $(1-p)$. The odds, which we will write as s, is the ratio

$$s \stackrel{\text{def}}{=} \frac{p}{1-p} \tag{25.2}$$

For the statement "the odds are 9 to 1", we infer that $p = 0.9$ so that the ratio of Equation 25.2 is 9.

More generally, when someone states that "the odds are 1 in n", we understand that the likelihood of some outcome is an observation $\underline{x}$ has the label +1 is $1/n$ of the likelihood that $\underline{x}$ has the label -1. That is, we are given the ratio of the odds as $s = 1/n$ and we want to find the probability that corresponds to the odds that someone has provided.

We can convert the odds to a probability by inverting Equation 25.2 so that, given s, we can solve for p. One solution is

$$\begin{aligned}
& s &=& \; p/(1-p) \\
\equiv \quad & s(1-p) &=& \; p \\
\equiv \quad & s - sp &=& \; p \\
\equiv \quad & s &=& \; p + sp \\
\equiv \quad & s &=& \; p(1+s) \\
\equiv \quad & p &=& \; s/(1+s)
\end{aligned} \tag{25.3}$$

We can now return to our problem in binary classification of data, which is: what is the probability that an observation $\underline{x}$ has the label +1? One method is to relate the scalar d – which is the signed distance from an observation $\underline{x}$ to the separating hyperplane $\mathbb{H}$ – to the scalar value s.

© R E Ellis 2025

We can constrain this relation by considering some necessary conditions that the relation must satisfy. Some of the properties we know or desire are:

- $d(\underline{x}; \vec{n}, c)$ is unbounded, that is, $d \in (-\infty, +\infty)$
- $s \in (0, +\infty)$ because:
  - s is positive, that is, $s > 0$, because $\lim_{p \to 0} s = 0$
  - s can have *any* positive value, because $\lim_{p \to 1} s = +\infty$
- the map $s \mapsto d$ must be invertible so that we can find $d \mapsto s$

There are infinite choices for a strictly monotonic function that satisfies all of these constraints. The constraints are illustrated in Figure 25.1.

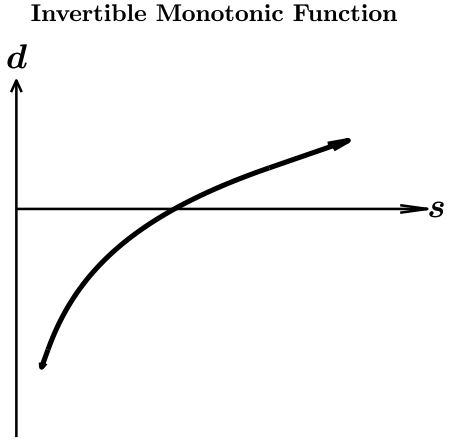

Figure 25.1: Constraints on a map from $s \in (0, +\infty)$ to $d \in (-\infty, +\infty)$. The function must be invertible so it must be strictly monotonic.

One function that satisfies the constraints, which we know from prerequisite material, is the natural logarithm. It is easy to verify that, for a value $s \in (0, +\infty)$, the computation $\ln(s)$ satisfies all of the constraints.

We will choose, from the infinitude of possibilities, to relate the ratio of odds s to the signed distance d as

$$\begin{aligned} & d &=& \ln(s) \\ \equiv \quad & s &=& \exp(d) = e^d \end{aligned} \tag{25.4}$$

© R E Ellis 2025

If we substitute the definitions of Equation 25.1 and Equation 25.2 into Equation 25.4, we get

$$\begin{aligned} d &= \ln(s) \\ \equiv \quad \underline{x}\vec{n} + c &= \ln\left(\frac{p}{1-p}\right) \\ \equiv \quad e^{\underline{x}\vec{n}+c} &= \frac{p}{1-p} \end{aligned} \tag{25.5}$$

Solving Equation 25.5 for the probability p gives us

$$\begin{aligned} p(\underline{x};\vec{n},c) &= \frac{e^{\underline{x}\vec{n}+c}}{1+e^{\underline{x}\vec{n}+c}} \\ \text{or} \quad p(\underline{x};\vec{n},c) &= \frac{1}{1+e^{-(\underline{x}\vec{n}+c)}} \end{aligned} \tag{25.6}$$

Equation 25.6 is referred to as the *logistic* function. It provides us with a way to compute the probability that the observation $\underline{x}$ has the label +1 based on the hyperplane $\mathbb{H}$ that is specified by the unit weight observation $\vec{n}$ and the bias value c. The logistic function is illustrated in Figure 25.2.

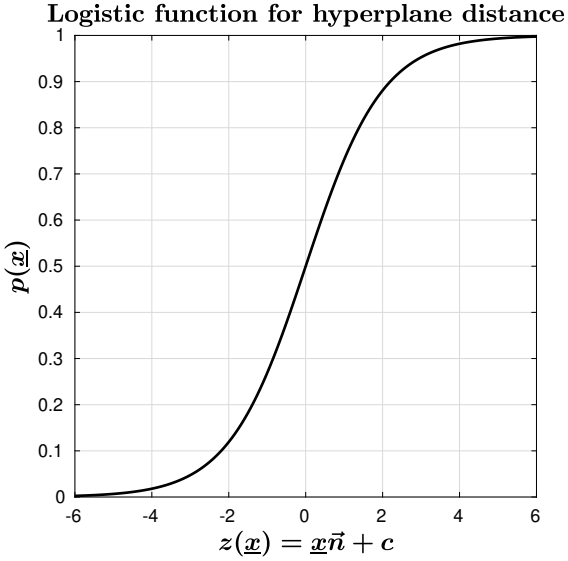

Figure 25.2: Logistic function for signed distance from an observation $\underline{x}$ to a hyperplane $\mathbb{H}$. The result is the probability $p(\underline{x})$ that $\underline{x}$ has the label +1 for a binary classification problem.

The logistic function is fundamental to many methods in current machine learning. A direct use is to perform regression not on the linear term $d = \underline{x}\vec{n} + c$, but on the logistic function $p(u)$. This concept of *logistic regression* can be used to understand maximum likelihood estimation and many other powerful methods for data analysis.

© R E Ellis 2025

Another use is in artificial neural networks. It is possible to arrange artificial neurons into *layers* that can learn a complicated classification task. The learning operates because the function $p(u)$ is continuous and differentiable, so the probability of classification can be propagated back from the output neuron to the neurons that provide its inputs.

Computational implementation of linear algebra has direct applications in data analysis. The closely allied concept of an observation space has allowed us to organize and approximate data observations of high dimension. Linear algebra is also fundamental to understand unsupervised clustering of data observations, classification of data observations by artificial neurons, and global optimization of classification by support vector machines. Its extension, by means of the logistic function, can be used to explore current methods in data analytics and machine learning.

25.1 Logistic Function – Some Properties

The logistic function is used widely in data analysis. Here, we can derive some simple properties. For this purpose, we will use the logistic function of a scalar argument d, which we can write as

$$p(d) = \frac{e^d}{1+e^d} = \frac{1}{1+e^{-d}} \tag{25.7}$$

From Equation 25.7, we can deduce a useful symmetry property:

$$\begin{aligned} 1 - p(d) &= \frac{1+e^d}{1+e^d} - \frac{e^d}{1+e^d} \\ &= \frac{1}{1+e^d} \\ &= p(-d) \end{aligned} \tag{25.8}$$

The derivative of the logistic function can be derived by recalling the Quotient Rule from basic differential calculus. Using the scalar argument d, we can write this rule as

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2} \tag{25.9}$$

© R E Ellis 2025

Applying Equation 25.9 to Equation 25.7, the derivative of the logistic function is

$$\begin{aligned} p'(d) &= \frac{e^d(1+e^d) - e^d e^d}{(1+e^d)^2} \\ &= \frac{e^d + e^d e^d - e^d e^d}{(1+e^d)^2} \\ &= \frac{e^d}{1+e^d}\,\frac{1}{1+e^d} \\ &= p(d)p(-d) \\ &= p(d)(1-p(d)) \end{aligned} \tag{25.10}$$

Equation 25.10 provides us with a way to compute the derivative of the logistic function that requires only a single invocation of the logistic computation.

25.2 Hyperplane Classification and Pseudo-Distance

Recall that, for a hyperplane $\mathbb{H}$, we have used two ways to specify $\mathbb{H}$:

a unit vector $\vec{n}$ and a bias scalar c *or* a non-zero vector $\vec{m}$ and a bias scalar b

A data observation $\underline{x}$ can be classified as being on the positive side of $\mathbb{H}$ if and only if:

$$\underline{x}\vec{n} + c \geq 0 \qquad \textit{or} \qquad \underline{x}\vec{m} + b \geq 0$$

In machine learning, we sometimes prefer the second specification. To do this, we introduce a *pseudo-distance* computation that is

$$u(\underline{x}; \vec{m}, b) \stackrel{\text{def}}{=} \underline{x}\vec{m} + b \tag{25.11}$$

For an observation $\underline{x}$ and a hyperplane $\mathbb{H}$, the distance d and the pseudo-distance u are related as

$$d(\underline{x}; \vec{n}, c) = \frac{u(\underline{x}; \vec{m}, b)}{\|\vec{m}\|} \tag{25.12}$$

That is, the pseudo-distance u and the distance d are related by a positive scalar. Either the sign of the pseudo-distance or the sign of the distance can be used to classify an observation $\underline{x}$, depending on how the hyperplane $\mathbb{H}$ is specified in a particular application.

We must take care when we use a non-zero normal vector $\vec{m}$ because the probability of an observation $\underline{x}$ occurring on the positive side of $\mathbb{H}$ is defined only for a unit vector $\vec{n}$. Using an incorrect specification of $\mathbb{H}$ can produce misleading results.

© R E Ellis 2025