ECE 469/ECE 568 - Machine Learning

October 1, 2024

Introduction to Classification in ML:

Input variables, Features or Attributes				
Tumor size	Texture	Perimeter	Outcome	Time
14.2	113	13.65	N	34
15.4	117	92.50	N	39
16.1	122	33.33	R	40
15.0	111	8.99	N	65

N= Non re-occurrence and R= Re-occurrence

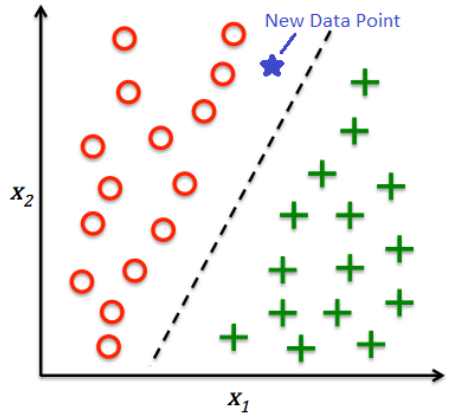
- The problem of predicting time is a numerical regression problem. We have already discussed regression in machine learning in detail.
- The problem of predicting re-occurrence/non re-occurrence is a (binary) classification problem.
- In this section, we are going to discuss classification for ML in detail.

Classification - Problem formulation

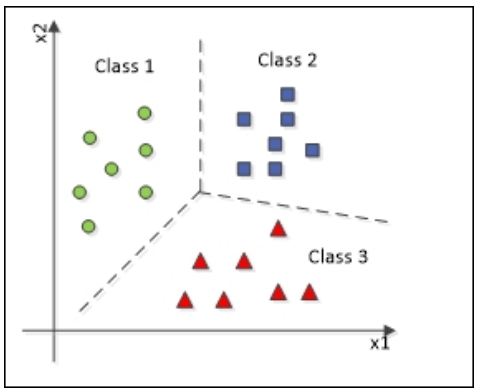
- We will be given a set of M training objects, $\mathbf{x}_1, \cdots, \mathbf{x}_N$. Each $\mathbf{x}_m$ for $m \in \{1, \cdots, M\}$ is a vector with a dimension L.
- For each object, we have been given a label y_m that will describe which class object m belongs to. Typically, this label is integer-valued.
- For instance, if there are only two classes in the data-set, $y_m\{0, 1\}$ or $y_m \in \{-1, +1\}$.
- More generally, if there are K classes, then $y_m = \{1, \cdots, K\}$.
- Our task is to predict the class y_{new} for an unseen object x_{new}.

Classification

- A binary classification problem is depicted below.



A multi-class classification problem is depicted below.



Types of classifiers

- There are two types of classifiers. They are
  - Probabilistic classifiers
  - Non-probabilistic classifiers
- In general, probabilistic and non-probabilistic classifiers differ in the type of output that the learning technique produce.
- In the probabilistic classifiers, the output is the probability of a new object belonging to a particular class: $P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y})$. This output probability is useful as it provides a level of confidence in the predicted output.
- In non-probabilistic classifiers, output is an assignment of an object to a class $y_{\text{new}} = k$, where $k \in \{1, \cdots, K\}$ denotes the available K classes.

Probabilistic classifiers

- In a probabilistic classifier, the output is the probability of a new object belonging to a particular class denoted by

$$P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}),$$

  where the training data is given in matrix/vector form $(\mathbf{X}, \mathbf{y})$.

- Thus, the output pf a probabilistic classifier must satisfy

$$0 \leq P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}) \leq 1$$

$$\sum_{k=1}^{K} P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}) = 1$$

- The output probability is directly related to a level of confidence in the predicted output.

Probabilistic classifier: Disease diagnosis applications

For instance, we consider a disease diagnosis application with two classes, healthy $(y = 0)$ and diseased $(y = 1)$. Providing the probability $P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y})$ is much more informative than simply stating that $y_{\text{new}} = 1$.

The Bayes classifier

- Given a set of training points from K classes, the task of the Bayes classifier is to be able to compute the predictive probabilities, $P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y})$ for each of K potential classes.
- Then these probabilities can be used for decision-making process, i.e., assigning $\mathbf{x}_{\text{new}}$ to the class with the highest probability.
- From Bayes' rule, the required predictive probability can be written as

$$P(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}) = \frac{f(\mathbf{x}_{\text{new}}|y_{\text{new}} = k, \mathbf{X}, \mathbf{y})P(y_{\text{new}} = k|\mathbf{X}, \mathbf{y})}{f(\mathbf{x}_{\text{new}}|\mathbf{X}, \mathbf{y})}$$

- Next, the marginal likelihood, $P(\mathbf{x}_{\text{new}}|\mathbf{X}, \mathbf{y})$ can be expanded to a sum over the K possible classes via the law of total probability as

$$f(\mathbf{x}_{\text{new}}|\mathbf{X}, \mathbf{y}) = \sum_{k'=1}^{K} f(\mathbf{x}_{\text{new}}|y_{\text{new}} = k', \mathbf{X}, \mathbf{y})P(y_{\text{new}} = k'|\mathbf{X}, \mathbf{y})$$

- The predictive probability can then be written as

$$f(y_{\text{new}} = k|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}) = \frac{f(\mathbf{x}_{\text{new}}|y_{\text{new}} = k, \mathbf{X}, \mathbf{y})P(y_{\text{new}} = k|\mathbf{X}, \mathbf{y})}{\sum_{k'=1}^{K} f(\mathbf{x}_{\text{new}}|y_{\text{new}} = k', \mathbf{X}, \mathbf{y})P(y_{\text{new}} = k'|\mathbf{X}, \mathbf{y})}$$

Classification - logistic regression

- In logistic regression, the output variable y can take a small number of discrete values.
- Thus, we now want to predict y just as we did in linear regression.
- First, we focus on a binary classification problem in which y can take on only two values, 0 and 1.
- For example, classification of emails as spam ($y = 0$) or legitimate ($y = 1$) is a binary classification problem.
- More examples: Tumor classification as malignant or benign and online transactions as fraudulent or legitimate.
- Typically, $y = 0$ is called the negative class, while $y = 1$ is called the positive class.
- If y takes more than two discrete values, then we have a multi-class classification problem. For example, $y \in \{1, 2, 3\}$ should have 3 classes.

Classification - logistic regression

- Our objective in classification is based upon an input vector $\mathbf{x}$, assign it to one of K discrete classes C_k for $k = 1, \cdots, K$.
- When $K = 2$, this is a binary classification. Whereas when $K > 2$, we have a multi-class classification problem.
- Typically, classes are taken to be disjoint.
- The input space is usually divided into decision regions.
- The boundaries of a decision region are called decision boundaries/surfaces.

Classification - logistic regression

- Recall that for regression problems, the output variable $\mathbf{y}$ was the vector of real numbers whose values we wish to predict.
- For two-class classification problem, we can use a binary representation in which there is a single output variable $u \in \{0, 1\}$ such that $y = 1$ represents class $\mathcal{C}_1$, whereas $y = 0$ represents class $\mathcal{C}_2$.
- Notably, we interpret the value of y as the probability that the class is $\mathcal{C}_1$. The corresponding probability values takes only the extreme values of 0 and 1.
- For $K > 2$ multiple classes, it is convenient to use a 1-of-K coding scheme, where we denote y to be a vector of length K.
- For example, for $K = 5$, a pattern from class 2 would be given the output vector:

$$\mathbf{y} = [0, 1, 0, 0, 0]^T$$

- Similarly, we interpret the value of y_k as the probability that the class is $\mathcal{C}_k$.

Classification - logistic regression

- Recall that we already motivated the logistic likelihood:

$$P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}_{\text{new}})}$$

- The above definition enables us to use our familiar linear model ($\mathbf{w}^T\mathbf{x}$) yet needed to transform it so that the output was a probability such that

$$0 \leq P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) \leq 1.$$

- Formally, the above logistic likelihood can be derived via log-odds ratio: $\ln(P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w})/P(y_{\text{new}} = 0|\mathbf{x}_{\text{new}}, \mathbf{w}))$.
- If $P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) \ll P(y_{\text{new}} = 0|\mathbf{x}_{\text{new}}, \mathbf{w})$, then this log-odds ratio will be a large negative value.
- If $P(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) \gg P(y_{\text{new}} = 0|\mathbf{x}_{\text{new}}, \mathbf{w})$, then this log-odds ratio will be a large positive value.
- Thus, the log-odd ratio can be modeled via familiar linear model: $\mathbf{w}^T\mathbf{x}_{\text{new}}$.

Classification - logistic regression

- The log-odd ratio is modeled via the linear model as

$$\ln\left(\frac{P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{P(y_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})}\right)=\mathbf{w}^T\mathbf{x}_{\text{new}}$$

- Moreover, notice that $P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})=1-P(y_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})$.
- Then we have

$$\begin{aligned}\ln\left(\frac{P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{1-P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}\right) &= \mathbf{w}^T\mathbf{x}_{\text{new}}\\ \frac{P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{1-P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})} &= \exp(\mathbf{w}^T\mathbf{x}_{\text{new}})\\ P(y_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w}) &= \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}_{\text{new}})}\end{aligned}$$

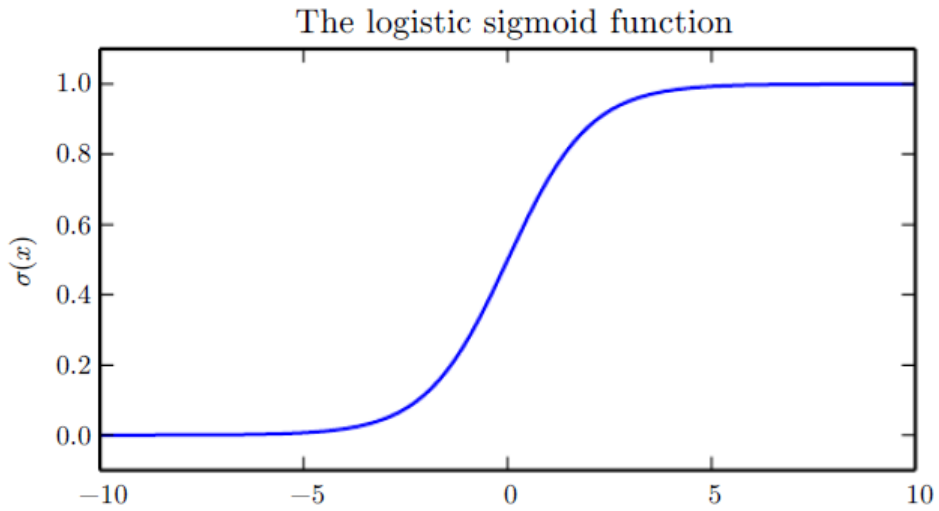
- Through this logistic likelihood, we actually model the log-odds ratio with a linear model.

Classification - logistic regression

- To facilitate this, we define a class of hypotheses $g(\mathbf{w}, \mathbf{x})$ as follows:

$$g(\mathbf{w}, \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})},$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is called the logistic function or the sigmoid function.



The sigmoid function is always bounded between 0 and 1.

Classification - logistic regression

- Thus our model is also bounded between 0 and 1.

$$0 \leq g(\mathbf{w}, \mathbf{x}) \leq 1$$

- Our model for logistic regressions is

$$g(\mathbf{w}, \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}.$$

- By using a training data set (x_i, y_i) for $i = 1, \cdots, M$, we can minimize a cost function to find optimal weights $\mathbf{w}$ for a given classification problem.

Classification - logistic regression

- Let us now consider properties of the logistic sigmoid function.

$$\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{\exp(x)+1} = \frac{\exp(x)}{\exp(x)+\exp(0)}$$

- The derivative of the logistic sigmoid function is given by

$$\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left(\frac{1}{1+\exp(-x)}\right) = \frac{\exp(-z)}{(1+\exp(-x))^2} \\
&= \left(\frac{1}{1+\exp(-x)}\right)\left(1-\frac{1}{1+\exp(-x)}\right) \\
&= \sigma(x)\left[1-\sigma(x)\right]
\end{aligned}$$

- The symmetry of logistic sigmoid function tells us that

$$1-\sigma(x) = \sigma(-x)$$

Classification - logistic regression

- We continue the properties of logistic sigmoid function as follows: By taking the logarithm of both sides of the definition, we have

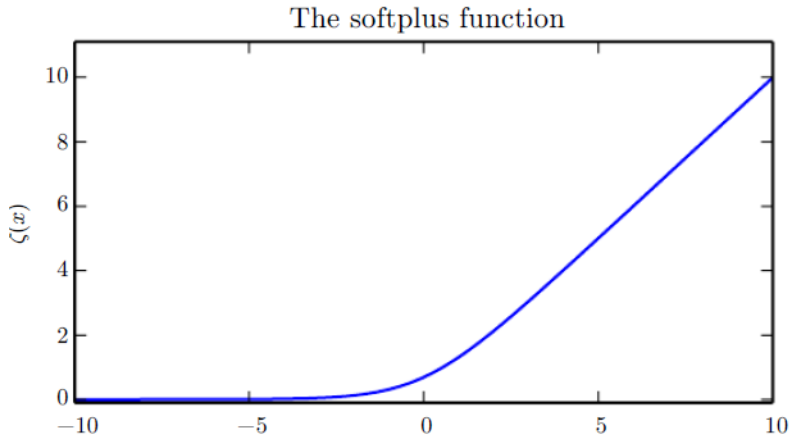
$$\ln(\sigma(x)) = \ln\left(\frac{1}{1+\exp(-x)}\right) = -\ln(1+\exp(-x))$$

- We define the softplus function as follows:

$$\zeta(x) = \ln(1+\exp(x))$$

- Softplus function is used as a activation function in ANN, and it can be viewed as a smooth version of ReLU activation.
- Then, the logarithm of the logistic sigmoid function can be written in terms of softplus function as

$$\ln(\sigma(x)) = -\zeta(-x)$$



Classification - logistic regression

- Since $\zeta(x) = \ln(1 + \exp(x))$, the derivative of softplus function is

$$\frac{d}{dx}\zeta(x) = \frac{\exp(x)}{1 + \exp(x)} = \sigma(x)$$

- Thus, the derivative of softplus function is logistic sigmoid function.
- Consequently, the integration of the logistic sigmoid function should be the softplus function:

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$$

- The inverse of the logistic sigmoid function is given by

$$\sigma^{-1}(x) = \ln\left(\frac{x}{1-x}\right)$$

- The inverse of the logistic sigmoid function is also known as the logit function.
- The inverse of the softplus function is given by

$$\zeta^{-1}(x) = \ln(\exp(x) - 1)$$