classification

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Classification is the task of predicting a discrete-valued label for a given data point, based solely on its features. The label belongs to a finite label space: binary classification uses two label values, and multi-class classification more than two. A hypothesis whose predictions take values in the finite label space is a classifier; it partitions the feature space into decision regions separated by the decision boundary. A widely used construction compares a real-valued hypothesis against a threshold to obtain the predicted label. A natural quality measure is the 0/1 loss; since it is neither convex nor differentiable, training methods minimize surrogates such as the logistic loss or the hinge loss, and the learned classifier is judged by the accuracy and the confusion matrix on a test set. When the label space is continuous, the task is regression.

Definition

A weather station records the minimum and the maximum air temperature of each day. From these two numbers, it must be decided whether the following day brings frost, that is, whether its minimum temperature falls below 0 °C. A library must file each incoming text into one of the categories math, novel, or engineering. Both are classification problems: the task of predicting a discrete-valued label y of a given data point, based solely on its feature vector $\mathbf{x}$, the vector of its features (Bishop, 2006; Duda et al., 2001; Hastie et al., 2009). Classification assigns each feature vector a label from a finite set. The label takes values in a finite label space $\mathcal{Y}$ and names the category of the data point. Binary classification uses a label space with two label values, such as $\mathcal{Y} = \{-1, 1\}$ with $y = 1$ marking a frost day; multi-class classification uses more than two label values, such as $\mathcal{Y} = \{\text{math}, \text{novel}, \text{engineering}\}$ for the texts. When the label space is continuous instead, the task is regression: predicting tomorrow's minimum temperature is regression, predicting whether tomorrow brings frost is classification. Classification is also distinct from clustering: as a form of supervised learning, classification learns from data points whose labels are known, while clustering groups data points without any given labels.

A hypothesis whose predictions take values in a finite label space is referred to as a classifier. A classifier partitions the feature space into decision regions, one region per label value, separated by the decision boundary (Fig. 1). Each classifier is fully determined by its decision regions: all feature vectors within the same decision region obtain the same predicted label.

A widely used construction of a classifier proceeds in two steps. First, a real-valued hypothesis $h : \mathcal{X} \to \mathbb{R}$ quantifies the confidence in one particular label value. Second, the confidence $h(\mathbf{x})$ is compared against a threshold to obtain the prediction $\widehat{y} \in \mathcal{Y}$. Logistic regression, for the label space $\mathcal{Y} = \{-1, 1\}$, uses a linear map $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ as the confidence in the label value 1 and compares it against the threshold 0,

$$\widehat{y} := \begin{cases} 1 & \text{if } h(\mathbf{x}) \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

The real-valued hypothesis h is not itself the classifier: the classifier is the thresholded map $\mathbf{x} \mapsto \widehat{y}$, and calling h the classifier is a slight abuse of language. The two decision regions are then the halfspaces on either side of the hyperplane $\mathbf{w}^\top \mathbf{x} = 0$.

A natural quality measure for a classifier is the 0/1 loss: its average over a dataset is the fraction of misclassified data points, and one minus that fraction is the accuracy. Even the best classifier cannot always achieve zero error: when data points with the same feature vector carry different labels, some misclassifications are unavoidable, and the smallest achievable risk is the Bayes risk. The 0/1 loss is rarely used as the objective function for training: it is neither convex nor differentiable as a function of the model parameters, so its minimization is computationally hard. Practical methods therefore minimize the average of a surrogate loss over the training set: a loss function that approximates the 0/1 loss while having more convenient properties, such as convexity or differentiability. Logistic regression uses the logistic loss, which is convex and differentiable (Bishop, 2006); the support vector machine (SVM) uses the hinge loss, which is convex but not differentiable (Boser et al., 1992; Cortes and Vapnik, 1995). Fig. 2 compares the three as functions of the margin $y \cdot h(\mathbf{x})$. The learned classifier is then judged by the accuracy and the confusion matrix on a test set.

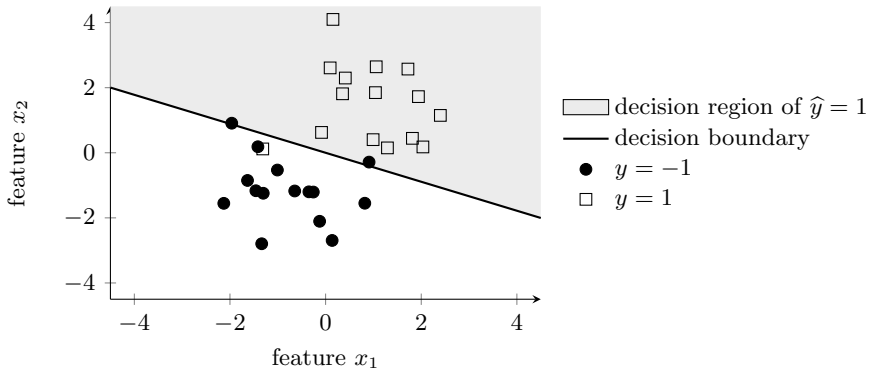

Figure 1: Binary classification with two features $x_1, x_2 \in \mathbb{R}$. Each marker is one data point of the training set, with its shape indicating the label. The line is the decision boundary of a linear classifier: it splits the feature space $\mathcal{X} = \mathbb{R}^2$ into two decision regions, and the predicted label of a data point is decided by the side on which its feature vector falls; the shaded region collects the feature vectors with $\widehat{y} = 1$. The two classes overlap, so even the learned linear classifier misclassifies some data points. Data generated by `pythondemos/classification.py`

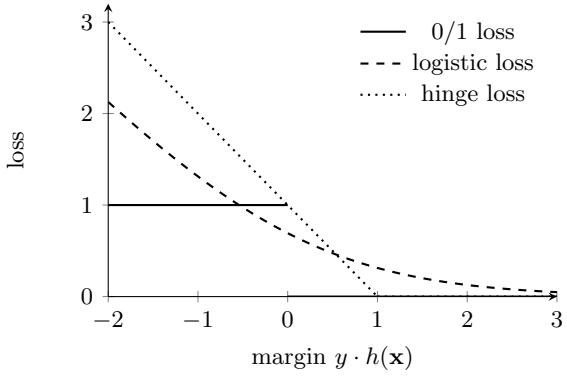

Figure 2: The 0/1 loss and two surrogates, drawn as functions of the margin $y \cdot h(\mathbf{x})$ for $\mathcal{Y} = \{-1, 1\}$. The logistic loss is convex and differentiable; the hinge loss is convex but not differentiable at margin 1

Cross References

label, label space, classifier, binary classification, decision boundary, decision region, 0/1 loss, logistic loss, hinge loss, logistic regression, SVM, accuracy, confusion matrix, regression, supervised learning, clustering, Bayes risk

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Boser et al. (1992). A Training Algorithm for Optimal Margin Classifiers. In: Proc. 5th Annual Workshop on Computational Learning Theory (COLT), pp. 144–152. ACM.

3. Duda et al. (2001). Pattern Classification. 2nd ed., Wiley, New York, NY, USA.

4. Cortes and Vapnik (1995). Support-vector networks. Machine learning, 20(3), pp. 273–297.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.