Multiclass SVMs

Disadvantages: using the decisions of the individual classifiers can lead to inconsistent results in which an input is assigned to multiple classes simultaneously. No appropriate scales. Training sets are imbalanced.

An important property of support vector machines is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum.

One advantage of SVMs is that, although the training involves nonlinear optimization, the objective function is convex, and so the solution of the optimization problem is relatively straightforward. The number of basis functions in the resulting models is generally much smaller than the number of training points, although it is often still relatively large and typically increases with the size of the training set.

A discriminant is a function that takes an input vector x and assigns it to one of K classes, denoted C_k.

In the previous chapter, we explored a variety of learning algorithms based on nonlinear kernels. One of the significant limitations of many such algorithms is that the kernel function $k(xn, xm)$ must be evaluated for all possible pairs xn and xm of training points, which can be computationally infeasible during training and can lead to excessive computation times when making predictions for new data points.

Least-squares solutions lack robustness to outliers, and this applies equally to the classification application

FLD

The projection onto one dimension leads to a considerable loss of information, and classes that are well separated in the original D-dimensional space may become strongly overlapping in one dimension. However, by adjusting the components of the weight vector w, we can select a projection that maximizes the class separation.

Although strictly it is not a discriminant but rather a specific choice of direction for projection of the data down to one dimension. However, the projected data can subsequently be used to construct a discriminant, by choosing a threshold y_0 so that we classify a new point.

Both the K-nearest-neighbour method, and the kernel density estimator, require the entire training data set to be stored, leading to expensive computation if the data set is large. This effect can be offset, at the expense of some additional one-off computation, by constructing tree-based search structures to allow (approximate) near neighbours to be found efficiently without doing an exhaustive search of the data set.

Basis Function

The simplest form of linear regression models are also linear functions of the input variables. However, we can obtain a much more useful class of functions by taking linear combinations of a fixed set of nonlinear functions of the input variables, known as basis functions. Such models are linear functions of the parameters, which gives them simple analytical properties, and yet can be nonlinear with respect to the input variables.

Ridge and Lasso

Regularization allows complex models to be trained on data sets of limited size without severe over-fitting, essentially by limiting the effective model complexity. However, the problem of determining the optimal model complexity is then shifted from one of finding the appropriate number of basis functions to one of determining a suitable value of the regularization coefficient .

Bias vs Variance

High variance can cause over fitting. High bias usually means less complex model. High variance usually means very complex model.

MLP or Neural Networks

An alternative approach is to fix the number of basis functions in advance but allow them to be adaptive, in other words to use parametric forms for the basis functions in which the parameter values are adapted during training. The most successful model of this type in the context of pattern recognition is the feed-forward neural network, also known as the multilayer perceptron.

SVM vs MLP

For many applications, the resulting model can be significantly more compact, and hence faster to evaluate, than a support vector machine having the same generalization performance. The price to be paid for this compactness, as with the relevance vector machine, is that the likelihood function, which forms the basis for network training, is no longer a convex function of the model parameters.

Hidden units are generally chosen by sigmoidal function.

For standard regression problems, the activation function is the identity so that $yk = ak$. Similarly, for multiple binary classification problems, each output unit activation is transformed using a logistic sigmoid function. For multiclass problems, a soft max activation function is used.

A key difference compared to the perceptron, however, is that the neural network uses continuous sigmoidal nonlinearities in the hidden units, whereas the perceptron uses step-function nonlinearities. This means that the neural network function is differentiable with respect to the network parameters, and this property will play a central role in network training.

Quiz Answers

The formula denoting the sample size needed to estimate the probability of a binary outcome using simple random sampling contains a term $z_{\alpha/2}$. What does α represent? **Ans** α represents the probability that you accept for your estimation being out of your accuracy margin, and the confidence level is equal to $1 - \alpha$.

Residuals

If the MLR model is well suited the residuals should be scattered symmetrically around the x-axis. If the residuals instead show some pattern (i.e. curve, funnel, etc.) you may need to consider a non-linear model or question the independence of your data points. Second, histogram of residuals can be a good plot for checking the assumption of normal distribution. Extreme outliers, for example, are easy to detect from a histogram and can bias a linear regression because they violate the normal distribution assumption.

SGD

State two important advantages and two disadvantages of using Stochastic Gradient Descent to determine model parameters. Advantage 1. Can handle large, streaming, possibly non-stationary datasets. 2. Stochastic nature of the gradient step can help the optimizer escape local minima. SGD is relatively simple to implement.

Disadvantage 1. The stochastic nature makes it hard to visualize/understand the update step for an individual data point. The cost function can go up for each such per-data point step (the overall cost function across all datapoints will still go down in general). 2. To get the best performance in terms of convergence speeds, the mini batch size has to be tuned. For the same reason, it may not be the best idea to run SGD on smaller datasets where simple GD would be sufficient. The performance difference is in typically terms of convergence speeds only, NOT accuracy, however SGD is sometimes able to escape local minima where GD would get trapped.

Lift vs ROC

Lift is appropriate when only the top few (most likely to be in the positive class according to your model) instances (e.g. people who will respond to a marketing campaign) need to be identified and one is not so concerned with the rest of the cases. If ROC is just used to provide AUROC (area under ROC), then it reflects performance over entire dataset, not only for the top few items. Even if the entire ROC curve is used, it is not easy to figure out the (multiplicative) factor of improvement over random model at any point in the curve, or how many samples that point corresponds to, so Lift is more interpretable and actionable even though ROC essentially contains the same info.

LDA assumes classes are normally distributed while FLD does not. If classes are normally distributed but covariances are quite different then use QDA.

Advantage of Naive Bayes: The probabilities of each class can computed in parallel.