Machine Learning Synopsis

Motivation: Why Study Machine Learning?

It adapts itself easily to solve various problems on its own, with little or no overhead from a programmer. This makes it a dream come true, which people had since they first though of building automata.

The Machine Learning Design Cycle

Preprocessing → Feature extraction / encoding → Feature selection → Machine learning → Model evaluation and selection → Postprocessing

LDA

Classify into classes using a line to divide them, this is called the hypothesis. It is possible to preprocess the data using so called basis functions. There is an analytic solution, though we need to invert S_W.

Linear Regression

Find the real number label for a novel data point. We fit a hyperplane to the existing points and if we get a new one it gets the corresponding value on this hyperplane. Has an analytic solution: (X.T X^-1) X.T y = w. Again we need to invert.

Mean of residuals = 0, variance = sigma^2, gaussian normally distributed

Logistic Regression

Find give the probability of a data point belonging to a class. This is sometimes better if the classes have different densities. Can't be solved analytically.

PCA

Projects the old features on new ones. It assumes that variance means relevance and finds new features based on that, by projecting the old features on the new ones. We can cut off new features with low variance.

ICA

We can unmix mixed sources based on the idea that the individual sources are less gaussian distributed than the mix of them. N-1 non gaussian distributed, linearly mixed sources, statistically independent.

Bias & Variance

Bias is how close is our average model to the perfect model of the hypothesis set.

Variance is how much do the models differ from data set to data set.

Validation methods

Hold out validation, k-fold cross validation, LOO, out of bootstrap validation, stratified cross val.

Regularisation

Early stopping, weight penalty, more data, sub- / oversampling, bagging, filter data.

Feature Selection

This is a way to discard useless features and only keep the good ones. The methods we learned about were forward feature selection, backward elimination and correlation analysis.

Auto ML

Auto ML describes the automatic hyperparameter selection for algorithms. Examples are bayesian optimisation, successive halving and hyperband approaches.

Bayesian Optimisation

It uses an acquisition function, which hopefully tells us the parameter settings, which tell us a lot about the high dimensional function of hyperparameters, that we want to optimise. It basically looks as the most informative points of the function, which are unknown. Then we try to minimise the loss using this info. Gets faster to the good results in the end, because it guesses the shape of the function well.

Hyperband

Pretty much like successive halving, but it only start very aggressively to reject models. Later it will not be as aggressive. Its pretty good in the beginning, because it loses all underperforming models extremely quickly.

SVM

A SVM is a classifier, which uses only the important points to create a hyperplane between the classes. This hyperplane is optimised to create the largest margin between the two classes, while still classifying the points well. There are hard-margin and soft-margin SVM. Soft margin SVM lets data points sit inside the margin, while hard-margin doesn't permit this.

Kernel

Kernel functions are functions, which calculate the inner product of a possibly higher dimensional space without first transporting the data vectors into that space, which saves us a lot of time.

SVM Regression

We want every data point in a tube around the hyperplane with the lowest margin possible. The tube, which has a diameter of 2 epsilon, is better if the margin is very small.

Kernel PCA

A PCA for nonlinear variances. We put it into a space, where the variance can be expressed by eigenvectors.

Trees

Trees are an easy and well interpretable method of solving classification and regression problems.
They split the data according to the feature with the most information gain. For regression trees we simply take the mean of the values as their leaf values. They overfit easily.

Forests

As single trees tend to overfit we combine multiple trees to achieve a lower variance.
Possible methods: Bagged trees, random forests, extra trees. Random forests use bootstrapping and a feature subset.

Boosting

Boosting is a sequential method to improve weak learners by fixing the errors of the last in the next part of the model.

AdaBoost

It iteratively creates new models by weighting the misclassified data points from before exponentially in their importance. Also the submodels are weighted by how well they perform.
Gradient Bossting

The more general version than AdaBoost, we can have any weak learner and also any loss function for our misclassified points, the weighting of the models also falls away.
We optimise our following model so that the new function fits the error residuals of the last. Using this and a shrinkage (learning rate) we basically use gradient descent to find good model weight without overfitting.

K-Means Clustering

We choose the number of classes as k, then initialise their centroids. Iteratively we sort all data points to their closest centroid and then choose the mean of each centroid as new centroid until convergence. Can't deal with weird shapes, variances, two clusters being together, or one cluster having two parts.

DBScan

We base our estimate of classes on them having points close to each other. We will have noise/outliers, core points and non core points. Core points are the ones with a minimum number of data points in a distance of epsilon (both hyperparameters). Non core points are the ones, which have a core point in their surrounding, but not enough neighbours to be a core point. Outliers are neither connected to a class, nor do they have enough points in their neighbourhood.

Independence & Separation

Independence is pretty much precision and separation pretty much recall.
Using the loan example for banks: independence is: For the loans given is it a 50/50 split? Separation in this example is: If the person would pay their loan back is it a 50/50 split in who gets the loan?