Be able to do all exercise, solutions are in brightspace

MACHINE LEARNING exam recap ( 5DAYS)

2016/17

Y Calculate F1 measure, precision, recall ++

Y Calculate the overall accuracy of the classifier across the 5 folds +

Y skewed class distributions problem when evaluating the

performance of a classifier, suitable evaluation measure for it

Y Calculate the overall entropy +++++

Y Using Information Gain, find the best feature to split the root of a Decision Tree classifier (or binary split) ++++

Y inconsistent data in the context of decision trees, how it is handled? +

--> in conflicting labes we use:

decision tree pruning ( class label associated with this new leaf node after e removal of a sub-tree with the root of that sub-tree becoming a leaf node)

Y Find nearest neighbour

Y performance metric for evaluation ++

Y contingency table of conditional and prior probabilities that would be used by Naïve Bayes

Y Naïve Bayes classifier +++

Y kNN classifiers ( and distance weighted) (best k choice) (predict exercise) (difference btw unweighted kNN and weighted kNN)+++

Y distance functions++++

Y Lazy / Eager learning approach in classification (example) +

Y curse of dimensionality as it applies to the kNN

Y Normalize numeric feature data ( important in KNN )+

#Correlation, covariance in linear regression

Y Cross validation ( difference with hold-out testing)+ +

#bias terms and variance in a neural network architecture++

Y Coefficent of determination +

2017/2018

Y diversity role in ensemble classification(bagging, boosting) ++++++++

Y Ensemble methods

Y what differentiates the different ensemble members in a Boosting Ensemble

Y Centroid vector of cluster in k-mean algorithm in dataset

Y k-Means clustering algorithm (stability problem) (scikitlearn) ( not possible with categorical data) ( when is not effective for finding clusters) ++++

Y difference between hyperparameters and other parameters in neural networks (example) ++

) Outline three practical reasons for reducing the number of dimensions of a dataset.

(ii) Describe how PCA performs dimensionality reduction. How might we go about se

lecting an appropriate number of dimensions for PCA? +

(i) Explain the difference between *filters* and *wrappers* for feature subset selection, with

reference to an example of each type of approach. +++

(ii) Why might filters and wrappers select different feature subsets when applied on the

same dataset? +

Explain how *backward elimination* works in the context of wrapper feature selection.

And forwar sequential search that is faster than backward elimination

T-test, p-value +

* Logistic or linear regression
* Gradient descend algorithm for regression model and update steps after every iteration ++ (and in ML generally)
* stochastic gradient descent optimisation difference from the standard gradient descent algorithm and difference to batch gradient descent ++

2018/2019

generalisation in the context of supervised learning, with refer

ence to the problem of overfitting.( causes) (scikit) (strsategies to prevent it) ++ overfitting and underfitting (in supervised learning)

example of how a cluster validation measure might be applied in unsupervised learning

Compute the equation of the least square linear regression

# key difference between agglomerative and divisive strategies for hierar

chical clustering

Explain why classification accuracy might not always be an adequate measure of predictive performance

2019/2020

Ranking classifier ( naive bayes) (application) (scikit lean functions for it)

ROC curve (example) (used for comparing btw classifiers)

real-world application of classification,

where the practical implications of a False Positive error and

a False Negative error might differ.

Silhouette method is a technique for quantifying the validity of a clustering

why training a multi-layer feedforward neural network is

considerably more difficult than training a single layer network+

Even in a simple single layer Feedforward Neural Network the units

(neurons) will have a fixed bias input. What is the reason for this bias

input? +

Explain in terms of the dynamics of a single neuron why it is that

single layer perceptrons are only able to learn patterns that are

linearly separable

Explain the operation of the following components in the training

of a neural network using gradient descent;

Cost function

Weight update

Stopping condition

2021/2022

binary classification exercise prior probabilities

Popular performance measures in regression are the R-squared

statistic, MAPE and MAE

In theory, to build a good classifier, the number of examples required

per feature increases exponentially with the number of features. In

practice, this is often not true. Give one reason why this might not be

true.

2022/2023 autumn

The Naive Bayes implementations in scikit-learn have a **fit-priors**

parameter that controls how the class priors are set.

Describe two strategies for setting the class priors and describe

scenarios where each of these scenarios would be appropriate.

Why is it important for Neural Networks to have cost (loss) functions

that are differentiable?

In developing *k-*Nearest Neighbour classifiers, what would be a

consequence of using a distance measure that is not a metric?

Confusion Matrices ( difference to Misclassification Rate)

2022/2023 spring

In *k-*Nearest Neighbour retrieval how can distances be calculated

for Ordinal features

. With scikit-learn, when StandardScalar normalisation (also known

as N(0,1) normalisation) is applied to data, what is the distribution

of the data after normalisation?

Decision tree, simpler is better

2 options for dealing with numeric (real valued) features

in a Naive Bayes classifier.

2023/2024 autumn

In *k-*Nearest Neighbour classification, it is common to set *k* to be

an odd number to avoid ties. Describe a situation where an odd

value of *k* will not always avoid ties.

b. With hold-out testing on a small dataset of fixed size, what is the

likely impact of increasing the size of the test set on the accuracy

estimate:

(a) increase the estimate and reduce the variance of this estimate.

(b) increase the estimate and increase the variance.

(c) decrease the estimate and reduce the variance.

(d) decrease the estimate and increase the variance

When training decision trees it is common to set the limit on the

number of samples required for a leaf node to control overfitting.

If this limit is reduced which of the following is most likely:

(a) training error reduced, test error increased

(b) training error reduced, test error reduced

(c) training error increased, test error increased

(d) training error increased, test error reduced

It is normal for Neural Network implementations such as

scikit-learn to use an alpha parameter for regularisation. What

does this do and how does this work?

e. Explain how mean absolute percentage error (MAPE) is

calculated. Outline one situation where it should not be used.

Initially, as new members are added to an ensemble the accuracy

improves. Eventually, the addition of additional members no

longer results in an increase in accuracy. Why is this?

measure for quantifying any bias in these

results. Calculate the overall bias.

It is standard practice to normalise the data in advance of fitting a

multivariate regression model but it is not essential.

If the data is not normalised, what impact will that have on

accuracy?

Describe one advantage of data normalisation in multivariate

regression.

The SGDRegressor implementation has a learning rate

parameter.

In terms of the stochastic gradient descent process, what

does this parameter control?

What are the consequences of setting a fixed (constant)

learning rate that is too large?

