机器学习考试准备

笔记本:	机器学习Pytorch		
创建时间:	2020/2/10 17:36	**更新时间:**	2020/2/16 23:32
作者:	li0121582@gmail.com		
URL:	https://www.google.com/search?sxsrf=ACYBGNQ1hvdLT8EnUvSx-8lcr6OhapD9e...		

机器学习考试准备

1.相关缩写

also:Residual Sum of Squares

一、SSE(和方差)

该统计参数计算的是拟合数据和原始数据对应点的误差的平方和，计算公式如下

$$SSE=\sum_{i=1}^{n} w_i(y_i-\hat{y}_i)^2$$

SSE越接近于0，说明模型选择和拟合更好，数据预测也越成功。接下来的MSE和RMSE因为和SSE是同出一宗，所以效果一样。

均方误差（MSE）

MSE （Mean Squared Error）叫做均方误差。看公式

$$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

这里的y是测试集上的。

Mean square error

均方根误差（RMSE）

RMSE（Root Mean Squard Error）均方根误差。

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}$$

这不就是MSE开个根号么。有意义么？其实实质是一样的。只不过用于数据更好的描述。
例如：要做房价预测，每平方是万元，我们预测结果也是万元。那么差值的平方单位应该是千万级别的。那我们不太好描述自己做的模型效果。于是干脆就开个根号就好了。我们误差的结果就跟我们数据是一个级别的，在描述模型的时候就说，我们模型的误差是多少万元。

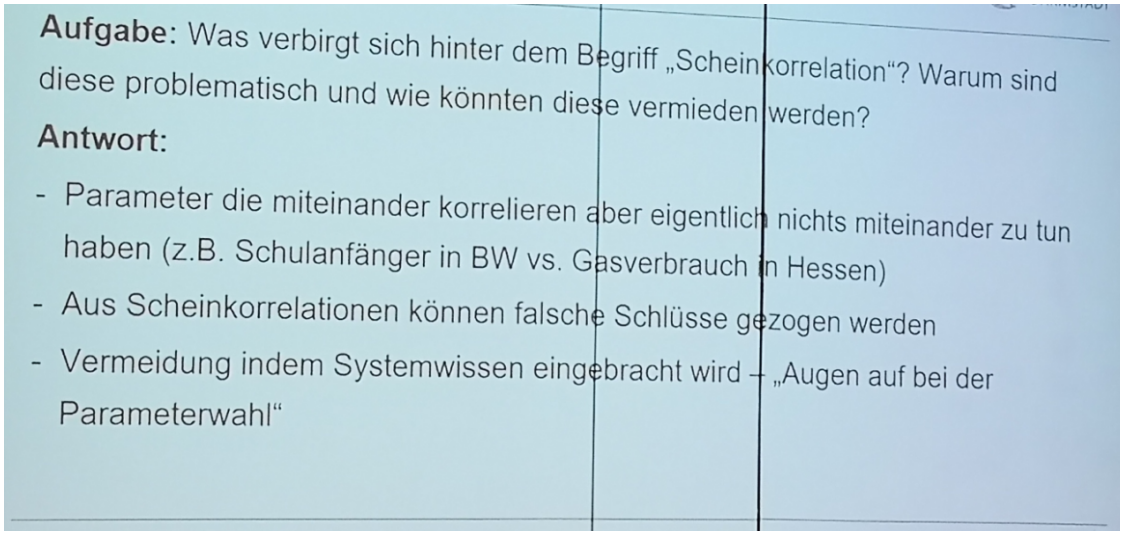
MAE

MAE(平均绝对误差)

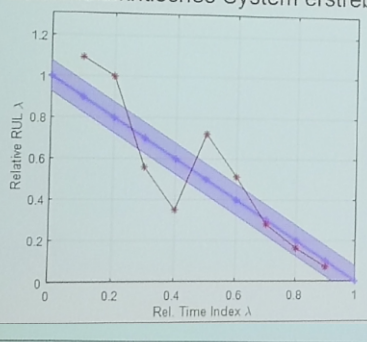
$$\frac{1}{m}\sum_{i=1}^{m}\left|(y_i-\hat{y}_i)\right|$$

Mean absolute error
RUL:remaining useful lifetime
PHM:Prognostics and Health Management
MBE:Mean bias error
RMS Value：Root Mean Square Value
(GPR)：Gaussian Process Regression
Data Mining Methodology for Engineering Applications (DMME)
RNN：recurrent neural network（循环神经网络）

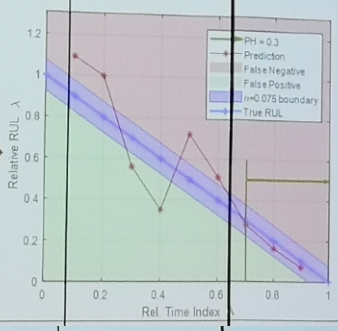
2.例题



Frage: Es wurden Prognosen zu unterschiedlichen relativen Zeitpunkten $\lambda = {}^{t_P}/_{t_{EoL}}$ durchgeführt. Die Ergebnisse eines Algorithmus sind in der folgenden Abbildung dargestellt. Lesen Sie den Prognostic Horizon ab und ermitteln Sie, wie viele Prognosen im „False Positive" und „False Negative" Bereich liegen. Welche Bedeutung hat „False Positive" und „False Negative" in diesem Kontext und welcher Bereich ist für ein sicherheitskritisches System erstrebenswerter?



Antwort

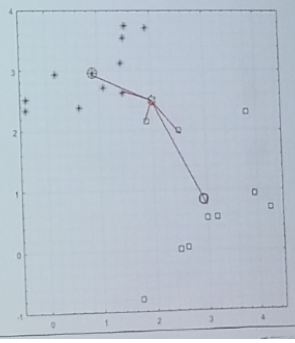


FN: 4x → frühzeitige Prognose
FP: 2x → zu späte Prognose
FP ist sicherer/ konservativer

Aufgabe: Im Folgendem soll ein „Neuer Punkt" einer Klasse zugeordnet werden. Hierfür soll zum einen der k-Nearest-Neighbor (kNN) und k-Means Algorithmus verwendet werden. Ordnen Sie den Punkt gemäß des **kNN** Algorithmus (**k = 3**) und gemäß des **k-Means** Algorithmus einer Klasse zu. Als Distanzmaß wird die **euklidische Distanz** verwendet. Beschreiben Sie kurz wie Sie jeweils vorgehen. Nehmen Sie Stellung zum Ergebnis der jeweiligen Klassenzuordnung und erklären Sie diese kurz. Welche Bedeutung hat das „k" in den jeweiligen Algorithmen?

Antwort:

- Klassifikation gemäß kNN bedeutet die **k** nächsten Punkte zu suchen → **Klasse** □
- Klassifikation gemäß k-Means bedeutet die Distanz zu den Klassenzentren zu bestimmen und das nächste zu nehmen → **Klasse** *
- Beide Algorithmen liefern unterschiedliche Ergebnisse. Dies liegt an der Vorgehensweise der Algorithmen
- Beim kNN gibt k die Anzahl der nächsten Punkte an, aus denen per Mehrheitsentscheid gewählt wird
- Beim k-Means repräsentiert k die Anzahl der Klassen – k spielt daher vor allem beim Clustering eine Rolle nicht bei der Klassifikation



3.PROCESS MODELS

1.主成分回归分析(principle component regression)

2.Overfitting

- **Overfitting problem:** → Too close to training data; does not generalize

  Starting position high bias: Reducing the bias causes the variance to go up which leads to an overfitting problem

- 交叉验证(cross-validation)
- 给评价函数加上正则项

and droupout

3.Underfitting

- **Underfitting problem:** → Too much generalized; training data not covered

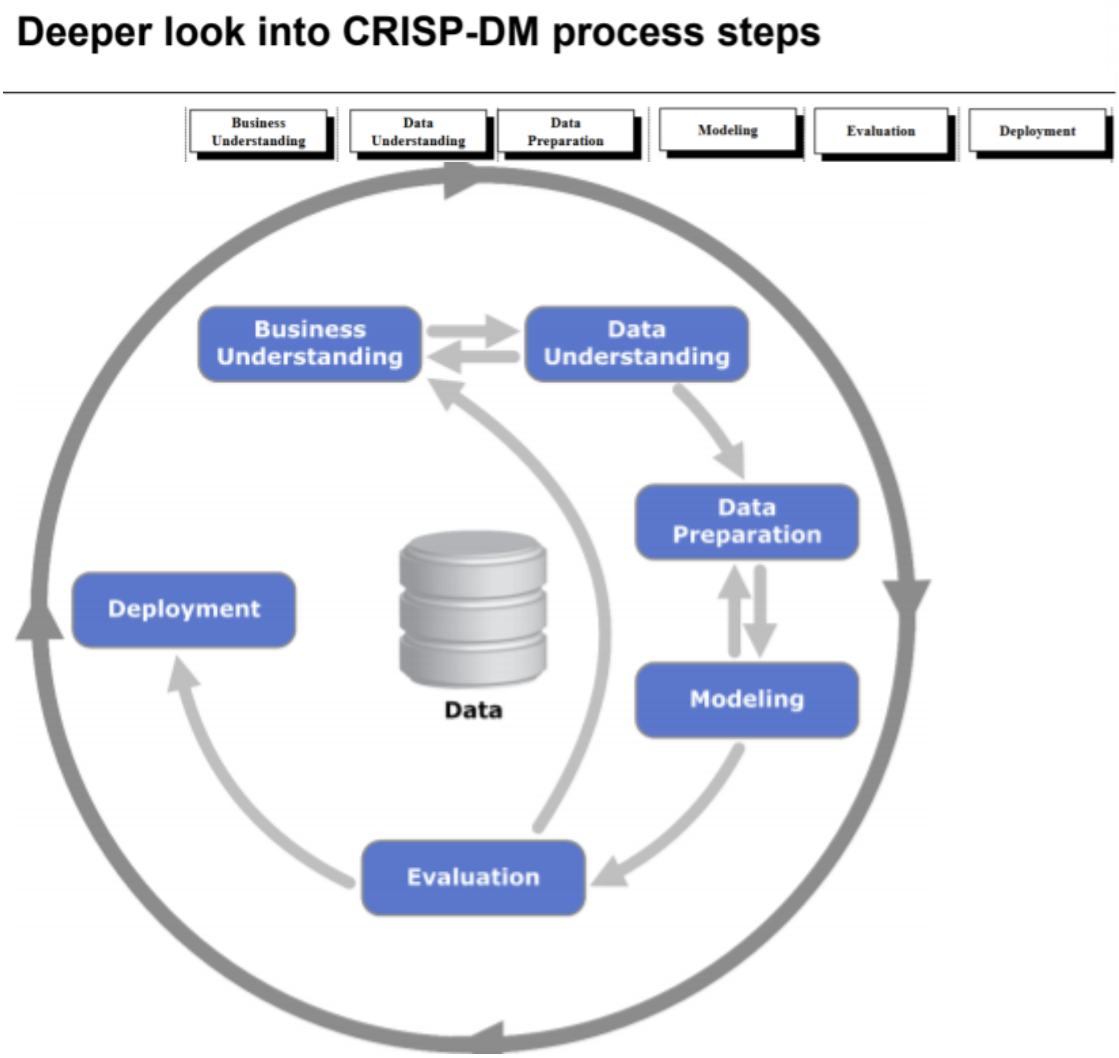
  Starting position high variance: Reducing the variance causes the bias to go up which leads to an underfitting problem

- 添加其他特征项
- 添加多项式特征
- 减少正则化参数

and more complex model

4.CRISP-DM:

Cross-Industry Standard Process for Data Mining



evaluation:

The CRISP-DM model – Evaluation
Error Metrics

Group	Metric	Derivation	Advantage	Disadvantage
Scale-dependant	Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$	+ Recommended for forecasting + High weight on large errors	- Sensitive to outliers
	Mean Absolute Error (MAE)	$MAE = \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \hat{y}_i\rvert$	+ Less sensitive to outliers than RMSE + Good to interpret	- Sensitive to outliers (less than RMSE)
	Median Absolute Error (MdAE)	$MdAE = \underset{i=1\ldots m}{median}(\lvert y_i - \hat{y}_i\rvert)$	+ Not very outlier-sensitive	- Harder to interpret than MAE and RMSE
Scale-independant	R^2-score	$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$	+ Standard metric in scikit-learn + Well-suited to estimate the generalization error + Normalized scale	- Sensitive to outliers
	Normalized RMSE (nRMSE)	$nRMSE = \frac{1}{n}RMSE$ with n = scaling factor	+ Normalized scale	- Sensitive to outliers - Scaling factor n has significant influence on the error metric

5.DMME Process

Data Mining Methodology for Engineering Applications

6.OSA-CBM

Open System Architecture for Condition-Based Maintenance

4.LINEAR MODELS AND EVALUATION

1.Curse of dimensionality(维数灾难)

Model complexity (p, N) and variance of estimates for different training datasets are directly related in linear models.

2.Confusion Matrix(混淆矩阵)

Confusion Matrix
(Classification)

TECHNISCHE UNIVERSITÄT DARMSTADT

	Classified as +	Classified as -	
Is +	true positive (tp)	false negative (fn)	tp + fn = P
Is -	false positive (fp)	true negative (tn)	fp + tn = N
	tp + fp	fn + tn	|E| = P + N

- The confusion matrix summarizes all important information
- How often is class *i* confused with class *j*
- Most evaluation measures can be computed from the confusion matrix
- Accuracy, Precision, Recall, Specificity, False Negative Rate, False Positive Rate

Frequently used are **Accuracy**, **Precision, Recall and Specificity**

$\text{Accuracy} = \frac{tp+tn}{tp+tn+fn+fp}$ $\text{Precision} = \frac{tp}{tp+fp}$ $\text{Recall} = \frac{tp}{tp+fn}$ $\text{Specificity} = \frac{tn}{tn+fp}$

精度 真阳性 真阴性率

5.TREE BASED METHODS & ENSEMBLES

1.decision tree

A decision tree is a tree structure in which each internal node represents a judgment on an attribute, each branch represents the output of a judgment result, and each leaf node represents a classification result.

ID3：

https://blog.csdn.net/alw_123/article/details/85116747

https://www.jianshu.com/p/b90a9ce05b28
信息增益等于信息熵减小
CART：
分类用Gini，越小纯度越高
回归用其他：SSE(sum of aquares)

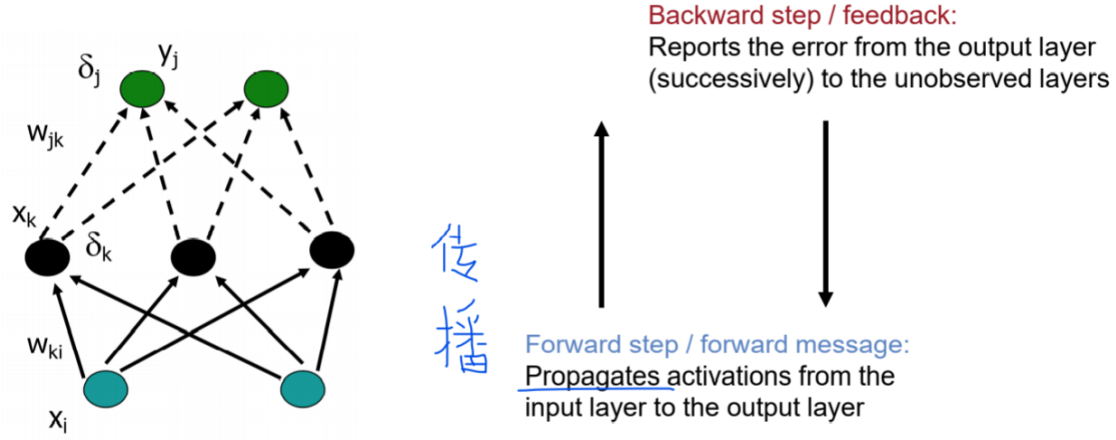
2.Pruning of Decision trees
Reducing Overfitting of the tree to the training data
Increase intelligibility（清晰度）

3.Random Forrests
A random forest consists of several uncorrelated decision trees. All decision trees have grown under a certain type of randomisation during the learning process.

https://www.cnblogs.com/yuluoxingkong/p/9386675.html

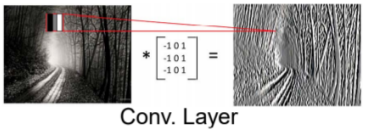
4.Clustering
Partitioning approaches
Hierarchical approaches

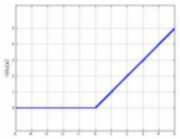
5.Backpropagation

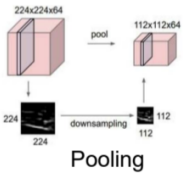


6.Deep Convolutional Neuronal Network (DCN/CNN)

- **Convolutional Layer**: Filters detect local patterns such as color values, edges,...
- **Rectified Linear Unit** (ReLU): Non-linear activation functions are applied per element
- **Pooling Layer**: Compress the representation (downsampling/sub-sampling). They are applied to each checkbox independently and are intended to make the network invariant to smaller transformations
- Output Layer with **Activation Function** Soft-Max







7.svm

SVM are based on statistical learning theory. They can be used for learning to predict future data. SVM are trained by solving a constrained quadratic optimization problem. SVM, implements mapping of inputs onto a high dimensional space using a set of nonlinear basis functions.

7.Self Organizing Maps / Best Matching Unit (自组织映射)

https://www.zhihu.com/search?type=content&q=自组织映射

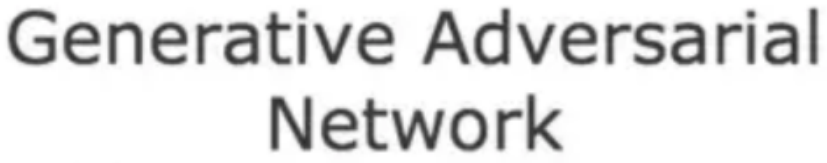
8.Generative Model

监督学习的任务就是学习一个模型（或者得到一个目标函数）,应用这一模型，对给定的输入预测相应的输出。这一模型的一般形式为一个决策函数Y=f(X)，或者条件概率分布P(Y|X)。监督学习方法又可以分为生成方法(generative approach)和判别方法(discriminative approach)。所学到的模型分别为生成模型(generative model)和判别模型(discriminative model)。判别模型求的是P(Y|X)，即后验概率；而生成模型最后求的是P(X,Y)，即联合概率

9.Accuracy Paradox

Accuracy Paradox for Predictive Analytics states that Predictive Models with a given level of Accuracy may have greater Predictive Power than Models with higher Accuracy.

10.GAN



11.Data Mining

Apriori Algorithm（关联规则）：

关联规则是指从一份资料库中（如销售记录）中发现某些特征（如商品种类）之间的联系。

12.OSA-CBM

Open System Architecture for Condition Based Maintenance

13.RTF

Run to failure is a maintenance strategy where maintenance is only performed when equipment has failed.

14.cumulative distribution function (CDF)

→How to calculate the RUL?

- → Mean of CDF reflects the expectation value of discrete distribution
- → Median of CDF reflects a probability of 50 % that a component will have failed until that time
- → Specify a distinct probability value for the CDF

15.training, validation and testing

- **Training:** the data that is used to train an algorithm (e.g. Neural Network)
- **Validation:** the data that is used to optimize the parameters
- **Test:** the data that is used to test the trained model – never seen by the algorithm before

16.Fault Tree Analysis (FTA):

Engineering tool to model failure dependencies for multicomponent systems based on a tree structure with the following characteristics:

17.Markov decision process, MDP

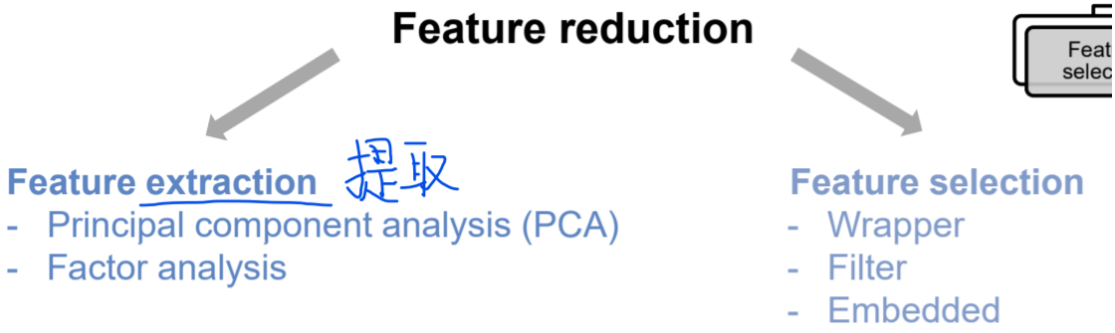
18.q-learning

状态（state）的价值(value)用v表示，（状态，动作）(state,action)的价值（value）用q表示（Reinforcement Learning：an Introduction）里头就是这么记的。然后这个q就一直延续到了Q-learning里了。也即Q值表示状态-动作对的值

6.DATA UNDERSTANDING & PREPROCESSING

1.Feature Engineering

- Avoidance of multi collinearities and redundant parameters
- Better generalizability
- Evalutaion of reduction methods through model performance/quality



7.单词

Maintenance
metric
pseudo
threshold
Accuracy Paradox
generative adversarial network
Apriori Algorithm（关联规则）
cumulative distribution function（累积分布函数）
degradation
algorithm
generative adversarial network
Advisory Generation
extraction
FMU (Functional Mock-up Unit)
Dymola