Project:

Otto product classification

It is to study classification methods.

1/ decision tree

2/ naïve Bayesian

3/NN

4/ logistic regression

5/ SVM

1/19/2018

Complex business problems/actionable insights/trends and patterns/metrics

Prepare for interview.

Data science:

* Collect , extract,query,clean,and aggregate data for analysis.
* Perform visual and statistical analysis on data using python and its associated libraries and tools.
* Build,implement, and evaluate data science problems using appropriate machine learning models and algorithms.
* Use appropriate data visualization tools to communicate findings.
* Create clear and reproducible reports to stakeholders.
* Identify big data problems and understand how distributed systems and parallel computing tehchnologies are solving these challenges.
* Apply functions, modeling, and validatition problem-solving processes to datasets from various industries in order to provide insight into real-time problems and solutions.

Questions:

1/ where do you see yourself in five years?

2/

1/20/2018

1/12/2018

Statistics-notes

Chapter1 ,概率论基本概念

E 随机试验： 1、可在相同条件下重复地进行 2. 每次试验的可能结果不止一个，并且已经知道所有可能结果 3. 进行一次试验前不可能知道结果

S 样本空间，E 的每个结果 为样本点

随机事件 由样本点组合而成。

频率， 概率

A1, A2,….是两两不相容的事件 A1A2=Ø

P（Ø） =0

1.等可能概率 1、 样本空间是有限个元素 2. 每个基本事件发生的可能性相同。

基本事件是两两不相容的

P(A)=A/S

2.条件概率

P(b/a)=p(ab)/p(a)

P(abc)=p(c/ab)p(b/a)p(a)

3.独立事件

P(ab)=p(a)\*p(b)

Chapter2 随机变量及其分布

1. 离散型随机变量分布
2. 连续型随机变量分布

Chapter 3的数学特征

数学期望，方差，相关系数，矩。

方差：研究随机变量与其均值的偏离程度。

Chapter 6 样本及抽样分布

直方图，箱线图

Markov model –

1/26/2018

Q-learning is a model-free reinforcement learning technique.

Linear-SVM

1/28/2018

Advanced knowledge of statistical techniques and machine learning.

Visualization tools-tableau, R Shiny

Skill to communicate complex ideas effectively.

Responsibilities

You will analyze and evaluate a wide range of data, including data from human resources information systems and relevant internal and external survey assessments.

Leveraging your advanced data analysis skills, you will create innovative approaches to answer our clients’ most relevant talent questions. You will be working on a variety of problems across many different industries, such as a top-tier professional services company trying to improve the efficiency of their recruiting pipeline, an insurance company facing high, inexplicable attrition, and a major pharmaceutical company in need of a roadmap to better understand their future talent needs.  
  
You will work with clients and our team to translate clients’ questions into analytical problems, manage the entire data cleaning pipeline, apply machine learning algorithms and data visualization techniques, generate insights, and communicate them back to clients.   Working on projects and exchanging experiences with your colleagues means you will face new intellectual challenges on a daily basis, while continuously building your methodological knowledge and skills.

Project:

ML project: house price prediction

( Understand problems – hypothesis generation – get data – data exploration – data pre-processing – feature engineering – model training – model evaluation )

Git- add an existing project to github using command line.

Notice that : do not initialize the new repo with README

1/ change the current working directory to your local project

2/ git init

3/ git add .

4/ git commit -m “ first commit”

5/ git remote add origin ~~remote repo url~~

6/ git push origin master