Programação

- Para R ou Python:
- fluência na sintaxe básica
- métodos e pacotes para leitura e escrita de dados (pandas)
- fluência nos principais pacotes de machine learning, como:
  - Python: sklearn
  - R: caret, mlr

Estatística Básica

- Variáveis Aleatórias Contínuas e Discretas
- Função densidade de probabilidade e distribuição acumulada
- Propriedades de distribuições: médias, medianas, quartis, moda, variância, etc.
- Testes de Hipótese
- Principais distribuições: Normal, Bernoulli, Binomial, Uniforme, Poisson, Geométrica

Álgebra

- Matrizes e Vetores
- Álgebra Matricial
- Distâncias e Produto Interno

Avaliação de modelos:

- métricas de avalição de modelo: KS, Gini, AUC, RMSE, MAE, F1, Recall, Precision, R2
- validação holdout, leave one out, k-fold, out of sample, out of time

Data Prep

- tratamento de missings
- tratamento de outliers
- categorização de variáveis contínuas e discretas
- PCA
- correlação / associação entre dados contínuos e entre dados discretos

Agrupamento

- kmeans / k-medoids
- dbscan
- algoritmos hierárquicos
- estratégias de definição do número de clusters (joelho/elbow, silhueta, distância intra-cluster, etc.)
- Gaussian Mixture Models

Classificação

- regressão logística
- naive bayes (Gaussiano x Bernouli)
- Knn
- árvore de classificação
- Random Forest
- Estratégias de Boosting: Gradient Boosting, ADA Boosting, etc.
- Redes Neurais
- SVM

Regressão

- regressão linear
- regularização L1 e L2
- árvore de regressão
- análise de resíduos

Banco de Dados

- modelo de banco de dados relacional
- sintaxe de SQL
- algebra relacional: Join, group by, order by e etc.
- chaves primárias, secundárias e estrangeiras

Outros (conhecimentos básicos):

- Hadoop e Hive
- Spark e Pyspark
- redes complexas e teoria de grafos
- ensemble modeling
- análise de séries temporais
- detecção de anomalia
- text mining
- deep learning e tensor flow
- reconhecimento de imagens
- speech analytics