Prediktiv analys

FÖRELÄSNING 11

Dagens fråga

- Vilket ämne skulle du kunna prata om i timmar?

Dagens agenda

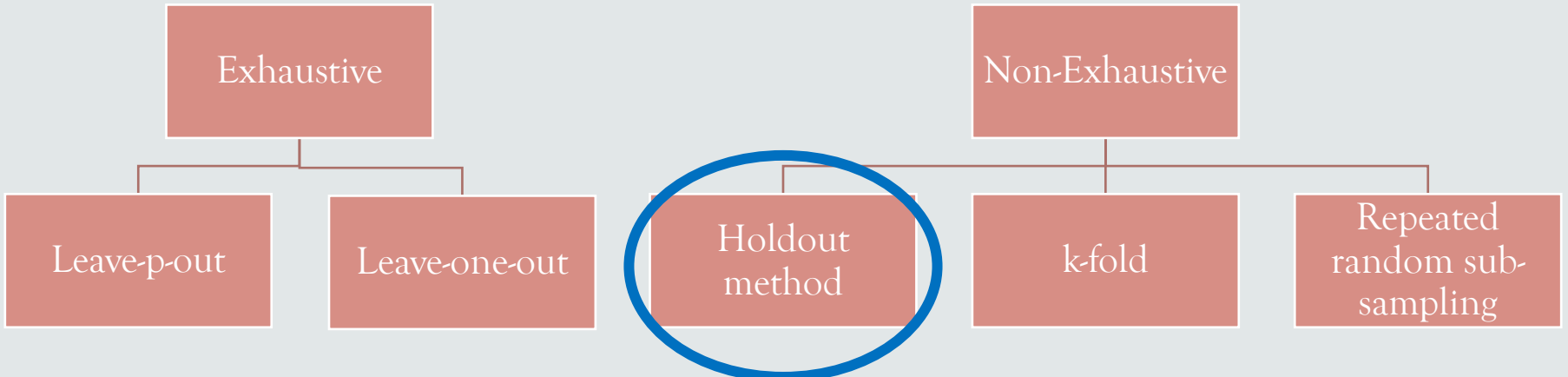
- Optimering av prediktionsmodeller
- Evaluering av modell: K-fold Cross Validation
- Hyperparameter tuning – hur välja bäst parameter till modellen?
- Exhaustive Grid Search
- Feature Selection

Förra föreläsning

- Optimering av prediktionsmodeller
- Ensamble methods
  - Random forest
  - Bagging
  - Bossting och AdaBoost
- Natural Langugae Processing NLP

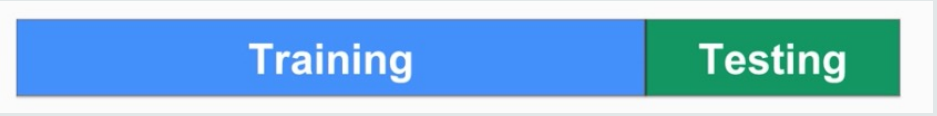
Hur man evaluerar (föreläsning 5)

- Målet med prediktiv analys är att få predikteringar om okända händelser
- Vi vill ha modeller som är generella mot den data som modellen inte sett tidigare
- För att estimera hur vår modell kommer prestera med data den inte sett tidigare använder vi en teknik kallad ”cross-validation”
- För att undvika overfitting



Cross-validation

- Målet med prediktiv analys är att prediktera okända händelser
- Skapa modeller generaliserar bra på osedd data
- För att uppskatta hur bra modellen presterar på ny data använder vi **cross-validation**
- Vi har använt **hold out** metoden med att dela datan i ett träning och test dataset
- Problemet är att den skapar enbart **ett** estimat av error metric till modellen
- For att en del modeller använder randomness kan vi få olikt resultat på grund av slumpen



K-fold cross-validation

- Datasetet delas in i K lika stora delar
- Turas om vilka dataset som används för träning och testning och modellen predikterar K gånger

K-fold cross-validation

- Efter K uppskattninger av vad error metrics är ges genomsnitten av dessa. Detta ger en bättre bild av prestandan till modellen
- Används inte bara för modellutvärdering men också för hyperparameter tuning (justering)
- Vanliga värden på K är 5 och 10
- Finns flera varianter som Leave One Out, Repeat K-fold
- Använda K-fold cross validation för att jämföra modeller

K-FoldCrossValidation.ipynb

Hyperparameter tuning

- **Hyperparameter** – parameter som inte lärs av modellen men som vi anger värdet av till modellen.
- I scikit-learn är hyperparapeterna satt till så bra default värde som möjligt, men det är inte givet de passar bra till just ditt problem.

```
RandomForestClassifier()
n_estimators; max_depth;
max_features; min_samples_split; min_samples_leaf;
min_weight_fraction_leaf; max_leaf_nodes;
min_impurity_decrease; min_impurity_split.
```

Hyperparameter tuning – Exhaustive Grid Search

Exhaustive Grid Search - brute force approach

1. Testar alla kombinationer av hyperparametrar från en grid (rutnät) med parametervärden
2. För varje kombination av hyperparameter utvärderas modellen med K-fold cross validation och valde error metrics
3. Bästa kombination av hyperparameter är den som returnerar bäst error metric

Hyperparameter tuning – Exhaustive Grid Search

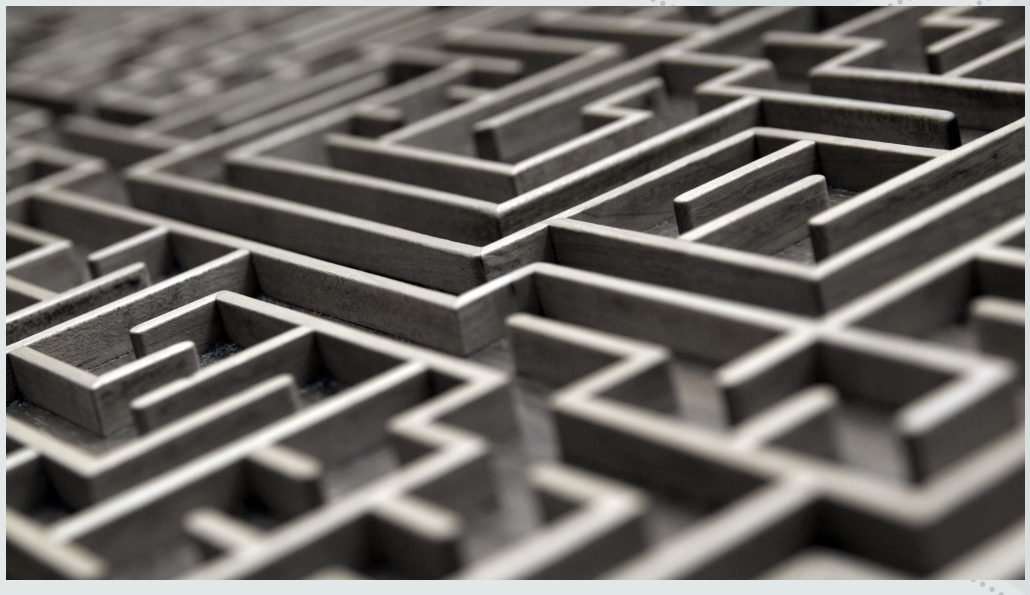
- **Parameter grid**

```
n_estimators=[10,30,50];
max_features=['auto','sqrt'];
max_depth=[5,10,20,30]
```

- Här $3 * 2 * 4 = 24$ hyperparameter kombinationer som kommer utvärderas
- För varje av de 24 modellerna kommer K-fold cross validation användas för utvärdering
- Om K=10 folds kommer datorn träna och evaluera 240 modeller
- Man kan välja vilka kombinationer man villa testa genom att ange grid med GridSearchCV
- **Använd aldrig hela datasetet för detta! Ha ett valideringsdataset för att testa hyperparameter**

Feature Selection

- Väldigt ofta är inte alla features relaterade med target så de hjälper inte till med prediktionen
- Att ha med irrelevanta features kan bidra till noise (brus) och ge bias (fel) till modellen
- Feature selection är tekniker för att välja de features som är mest relevanta och användbara för vår modell
- Det finns många metoder – bara några nämnas här



Feature Selection –
Ta bort dummy features med låg varians

- Dummy features med låg varians har är sannolik att ha väldigt låg påverkan på prediktionen
- *Ex: Tänk ett dataset med **kön** som feature och 98% av observationerna är **kvinna**. Det påverkar antagligen inte prediktionen för nästan alla datapunkter tillhör samma kategori, ingen varians.*
- Undersök sådana dummy features mer noggrant!
- Ta bort alla dummy features som är 1 eller 0 i mer än x% av datapunkterna.
- Eller skapa en minimum threshold för variansen till dummy features

$$Var[X] = p(1 - p)$$

där p är andel dummy features

Feature Selection –

Identifiera viktiga features statistisk

- Använda statistiska tester för att identifiera och välja relevanta features

För klassificering:

- **ANOVA F-statistik** för att evaluera förhållandet mellan numeriska features och target
- **Chi-squared test** för att evaluera förhållandet mellan dummy features och target

- I scikit-learn använder man SelectKBest

Feature Selection

-

Recursive Feature Elimination (RFE)

- **RFE** kan användas på modeller som beräknar **koefficienter** (linear och logic regression) eller modeller som beräknar **feature importance** (random forest)

1. Först väljer vi ett antal features vi vill använda i modellen
2. Modellen tränas på **alla** features
3. Sedan baserad på feature importance/koefficienterna blir de minst viktiga features eliminerad
4. Denna processen fortsätter till vald antal features är nått

Quiz frågor

Finns i Teams

Vad har vi gjort idag?

- Optimering av prediktionsmodeller
- Evaluering av modell: K-fold Cross Validation
- Hyperparameter tuning – hur välja bäst parameter till modellen?
- Exhaustive Grid Search
- Feature Selection

Nästa lektion

- Repetition
- Inlämning