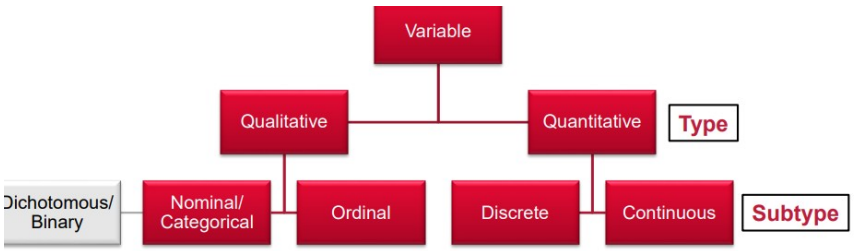
Klausurvorbereitung: **Datenanalyse**

Woche 1: Deskriptive statistik

Basic decriptive statistics

- **Population**: all elements whose characteristics are going to being studied
- **Sample**: a portion of the population selected for study
- **Variable**: a characteristic of the population we want to study
- **Elements of the dataset**: a specific subject included in the sample
- **Observation**: the value of a variable for an element



Variable type	Definition
Qualitative variable	
Nominal	Unordered nonnumeric categories
Ordinal	Ordered nonnumeric categories
Quantitative variable	
Discrete	Counts or certain non-intermediate values
Continuous	Any numerical value from a certain interval

Summary measures:
Based on location
Mode: The most common value in the dataset (there can be more than one mode, not all values are used)
Median: The value in the ‚middle' of an ordered sample, ie. 50% of the observations are lower than median and the rest 50% of the observations are larger
Mean: The ‚central' tendency of the data. (downside: The mean is affected by outlying values, not suitable for skewed data)

Dispersion measures:
Variance: It measures how much on average the observations are spread from the mean value of a continuous variable $\mathbf{s^2}$

Standard Deviation: It is preferred to variance because the variance is measured in squared units whereas the standard deviation is measured in the same units as the observations. Variance is always positive! A zero variance indicates that all the observations are identical!

Range it is defined as the difference between the maximum and minimum observation for a specific variable: Dispersion measures $R = maximum - minimum$

Interquartile range It is defined as the difference between the 75%th and 25%th observation (i.e., $Q3 - Q1$) within which the 50% of the observations are included:

Woche 3: Studiendesign
PICOT framework:
Patient, Intervention, Comparison, Outcome

Observational study types
- Longitudinal prospective/retroperspective cohort
- Case control study

Woche 4: Bivariate Analyse

Dependent variable	Independent variable		
	metric	binary	categorical
metric	regression analysis	two-sample t-test	one-way ANOVA
binary	regression analysis	Fisher's exact test	chi-squared test
categorical	regression analysis	chi-squared test	chi-squared test

Testing for normality:
- Are mean and median similar?
- Look at histogram and QQ plot (Quantile zweier statistischer Variablen gegeneinander abgetragen werden, um ihre Verteilungen zu vergleichen)
- Statistical test (shapiro wilk or kolomogrov (if n >2000)

If at least one group is not normal distributed, non-parametric test need to be applied

t- test:
- parametric test, variance does not differ i.e., a non-significant Levense test (test if variances are equal), also for paired samples
- if variance differs, apply Welchs t test

Mann-Whitney/Wilcoxon Rank sums
- non-parametric, independent groups,Wilcoxon signed-rank sums for paired samples

ANOVA
- more than two groups are compared regarding a metric variable
- uses between group and within group variation as measure if groups differ
- F statsistc: F>1 implies, that between group variation is higher than within group variation, so that at least one group has different population mean
- Assumptions: Homoscedacity (are variances equal?) test with Bareltt test. If significant Bareltts test, apply Welch ANOVA, then post hoc Tukey

Kruskal-Wallis
- No-parametric to ANOVA
- Ususally bunns posthoc test is applied

Pearson chi^2
- Pair of nominal variables
- Total number of units must be large enough
- Expected frequencies must be at least 5 (solution: merge some catergories to n>5 observations)

Fishers exact test:

- For two binary variables

Pearson, spearman and kendall tau correlations..

Effect measure: describes the association of two variables: outcome and intervention, it compares the outcome between two or more interventions

Effect size: describes the magnitude of effect measure, informs about clinical significance

For binary outcomes: risk and odds

Risk (p) to experience a certain outcome is **p = No of events in group/ total n in group**

- P ranging from 0 -100 %, he descriptive statistic of the outcome in each group

Odds

Odds = No events in 1 group / No of **no** events in 1 group (same as odds = p/1-p)

Risk ratio

- Is the ratio of the risk in group A to the risk in group B, **RR**= p_A /p_B
- The risk of the outcome decreases by a factor of X in intervention A as compared to B
- Reciprocal risk ratio:

RR > 1 : risk of group A is greater than in group B
RR = 0 : risk in group A equals to group B
RR <: risk of group A is lower than in group B

Odds ratio

- Is the ratio of the odds in group A to the risk in group B, **OR**= $odds_A$ / $odds_B$
- Reciprocal odds ratio:

OR > 1 : odds of group A is greater than in group B
OR= 0 : odds in group A equals to group B
OR<: odds of group A is lower than in group B

➔ Das gleich für logOR nur mit 0 anstatt 1

Log risk ratio standard error:

- Standard error measures the property of estimated parameter, how well it is estimated

Fomular:

Confidence intervals:

- A CI is a interval, that includes 95% of the parameter values if we repeat an experiment many times
- CI informs about stat. sign. of the value compared to the null value

- CI indicates weather clinically implausible values are observed

➔ OR and RR need to be ln transformed because they are not normally distributed

Absolute effect measures
Risk differences:
- the difference between the risk between group A and group B, RD = $p_A - p_B$
- Also has its own calculation of approximate standard error and CI

Mean difference and approx. standard error and CI
- MD: $x_a - x_b$
- two formulas for SE, either 1) assuming same SD in each group 2) assuming different SD in each group

Woche 5: sample size calculcation

Is the acquired sample **representative** of the target population?

Proper **sampling frame**: source from which the population units will be collected

Is the sample **large enough** to ensure accurate and precise results?
This requires
- knowledge of the available, relevant evidence
- a decision framework on how much error is allowed from collecting a fraction of the target population
- determine the minimum clinically important difference (MCD)

Sampling methods
- are divided into probability methods and non-probability methods

Probability sampling methods	Non-probability sampling methods
• Simple random sampling	• Convenience sampling
• Systematic random sampling	• Judgment sampling
• Stratified random sampling	• Quota sampling
	• Snowball sampling

Simple random sampling: each unit has the same probability of being selected, homogenous and small population
Systematic random sampling: population is initially ordered by a specific characteristic. Then, the sample is selected using a skip interval that has been pre-determined to achieve the required sample.

Stratified random sampling target population is initially divided into subgroups (known as *strata*) based on a characteristic (e.g., age or gender). Then, the units are sampled randomly from each stratum separately. The sample comprises all units sampled from each stratum.
Convenience sampling: relies upon convenience and access.
Judgment sampling: relies upon the belief that participants fit specific characteristics (based on the researcher's judgment).
Quota sampling: emphasises the representation of specific characteristics. Initially, the target population is divided into subgroups, similar to stratified sampling. Then the researchers use their judgment to select subjects or units from each segment based on a pre-specified number

Sample size calculation:
What we need to know:

- heterogeneity of target population, the higher the more sample we need
- expected MCID, from literature
- sampling error (usually at 5%)
- desired confidence in estimating effect size (usually 80%)

Type I error, alpha α.

- It refers to the probability of (falsely) rejecting the null hypothesis when it is true
- type I error indicates the probability of a '*false positive*' finding. It is typically specified at 5%

Type II error, beta β

- refers to the probability of (falsely) failing to reject the null hypothesis when it is false.
- type II error indicates the probability of a '*false negative*' finding. It is typically specified at 20%.

Power

- Power is the probability of (correctly) rejecting the null hypothesis when it is false. In short, power is one minus the type II error (1 – β), typically specified at 80%.

Relationship:
As the power increases, the total sample size increases as well. On the contrary, as the type I error increases, the sample size decreases exponentially.
Woche 6: Maschinelles Lernen

Ziel von ML:

- Generalisierbarkeit des Modells auf unabhängige Daten
- finde generalisierbare und prädiktive Muster in den Daten mit Hilfe von Trainingsdaten
- Muster: Zusammenhänge zwischen Variablen (z.B. Response und Prädiktor)

Supervidiertes Lernen:

- Trainingsschritt mit gelabelten Inputdaten
- bestimmte f(x), setzte x in f(x), berechne y gemäß von f(x)
- Beispiel: Regression, Klassifikation

Unsupervidiertes Lernen:

- Trainingsschritt mit **un**gelabelten Inputdaten
- Detektiere Ähnlichkeiten und Unähnlichkeiten in Daten
- Beispiel: PCA, hierarchisches Clustering, K-means clustering

Reinforcment Learing (Roboter)

Univariate lineare Regression:

- Responsevariable y z.B. Alter
- Prädiktorvariable x_a z.B Metabolit

y abgeschätzte (estimated) Responsvariable
β_a estimated Regressioncoefficient der Steigung von x_a
x_a Konzentrationswert des Metabolits
β_0 Achsenabschnitt

β_0 und β_a sind unbekannt und müssen mit den Trainingsdaten gelernt werden

Residuum= Differenz zwischen wahrem und vorhergesagtem Responsewert
Residual Sum of Squares = Summe der einzelnen Residuen zum Quadrat

Least Square Ansatz = bestimme Modellkoeffizienten so, dass RSS minimiert wird

Wie gut wurden Modelkoeffizienten geschätzt? Berechnung des standardfehlers SE

RSE: residual standard error ist Wurzel aus RSS/n-2
95% CI für β_a ist β_a +/- 2 mal die SE von β_a

Wie berechnet man die Signifikanz des Models?

- **Nullhypothese ist** β_a ist gleich 0, da in diesem Fall die Gleichung nicht lösbar ist
- **Alternativhypothese ist** β_a ist ungleich 0
- Also muss β_a weit genug von 0 weg sein, dies ist abhängig vom geschätzten β_a und den Standardfehler SE (β_a)
- Daraus kann man den t-Wert errechnen
- **p=Wert** bedeutet, dass ein Wert gleich t-wert oder extremer ist, unter der Annahme dass die Nullhypothese war ist

Wie gut beschreibt das Modell die Daten?

- Je kleiner der RSE (residual standard error) desto besser das Modell
- R2 bestimmt wieviel Anteil der Varianz durch die Daten erklärt wird

Formel R2

Multiple lineare regression:
- β_a beschreibt den mittleren Effekt auf y für einen one unit increase von x, während alle anderen Variablen fix gehalten werden

Wie berechnet man die Signifikanz des Models?

- Nullhypothese: alle $\beta_a = 0$
- F statisitk: berechtet aus den Total Sum of Square TSS und Residual Sum of Square RSS
- Verteilung von F-Statistik bestimmt den P-Wert

Welches sind die wichtigen Prädiktoren?
Varaibalenselektion über unterschiedliche Verfahren

a) Foward Selektion (füge Variable mit dem niedrigsten RSS zum Null-Modell hinzu)
b) Backward selection (entferne Variable mit höchsten P-Wert)
c) Mixed selection: Kombinationen aus beiden

Wie gut beschriebt das Modell die Daten?

- Berechne RSE (residual standard error) und R^2
- je mehr Variablen im Modell, umso höher R^2, daher nutze nur R^2 adjusted

Neues Datenset:

➔ setze Prädiktorenwert ein die Gleichung ein, berechne R2 und RSE auf unabhängigen Daten, beurteile grafisch die plots von y und estimated y

Was mache ich mit qualitativen Prädiktoren?

- Dummy variable erstellen (0,1 oder -1,1)

Formel:

Probleme der linearen Regression:

- Nicht lineare Zusammenhänge zwischen Prädikotren und Repsonse
- - check hierfür die Residuen, löse mit Datentransformationen
- Korrelation der Residuals: Varianz der Fehlerterme mit steigendem Responswert

Woche 7: Klassifikation

- Responsevariable ist kategorisch oder quantitativ
- Es wird die **Wahrscheinlichkeit** vorhergesagt, einer bestimmten Kategorie anzugehören
- Post hoc Klassifizierung basierend auf einer Entschiedungsregel

Logistische Regression vs lineare Regression

- lineare Regression klassifiziert Werte außerhalb von 0 und 1
- logistische regression sagt Wahrscheinlichkeit der Gruppenzugehörigkeit an, nicht direkt die Klasse

p(x) = Probability(x)

Formel der logistischen Regression:

Odds funktion:

➔ Odds ratio ist größer als 1 : wahrscheinlich Klasse 1 anzugehören

➔ Odds ratio ist kleiner as 1: wahscheinlichkeit Klasse 2 anzugehören

Fitten des Models:

- β_0 β_1 müssen bestimmt werden und müssen mit Trainingsdaten geschätzt werden
- **Maximum-Likelihood Methode:** bestimme β_0 β_1 so dass die vorhergesagte Wahrscheinlichkeit $p(x_i)$ der Klasse 1 anzugehören so nah wie möglich an der wahren Klasse des Individuums i liegt

Likelihood funktion

Vorhersage der Klassifikationsmodelle:

- P(x) kann für neue Werte berechnet werden und gibt einen xx % Wert aus
- Jetzt muss Entscheidungsregel bestimtm werden z.b > oder < 50% ist man Klasse y oder x

Evaluation der Klassifikationsperformance

Resampling

- Muss an unabhängigen Testdaten getestet werden (z.B. original Datensatz dritteln)
- Oder mit **resampling Methode**: hier wird mehrmals aus dem gesamten Datensatz Trainigs-und Testdaten gezogen um den Generalisierungsfehler zu verkleinern (Nachteil: Trainings und Testdaten sehr klein)
- Oder gleich andere Kohorte benutzen
- **Kreuzvalidierung** (5-fold crossvalidation): splitte Datenset k-mal, k-1 wird als Trainingset verwendet. Wiederhole k-mal das Modell und mittel Performancemaße
- **Bootstrapping:** ziehe randomisiert mit zurücklegen eine Anzahl an samples für Trainingsset, übrige Samples als Testdaten

Maße zur Klassifikationsperformance:

• positive (P): wahre Klasse ist positiv
• negative (N): wahre Klasse ist negativ
• predicted positive (PP): vorhergesagte Klasse ist positiv
• predicted negative (PN): vorhergesagte Klasse ist negativ
• true positive (TP): Modell sagt korrekt positive Klasse vorher
• true negative (TN): Modell sagt korrekt negative Klasse vorher
• false positive (FP): Modell sagt positive Klasse vorher, wobei die wahre Klasse die negative Klasse ist
• false negative (FN): Modell sagt negative Klasse vorher, wobei die wahre Klasse die positive Klasse ist

Accuracy (ACC) = TP + TN / P +N

- Funktioniert nicht gut bei unbalancierten Datensätzen in denen eine Klasse sehr selten ist

Precision = TP/PP (welcher Anteil an positiven Vorhersagen war wirklich korrekt?)

- z.B. Wert 0.5: wenn Modell Tumor ist bösartig vohersagt, ist dies zu 50% korrekt

Sensitivity (Trefferquote, recall, true positive rate TPR) = TP/TP+FN oder TP/P

- v.B. Wert 0.11: Modell identifiziert 11% aller bösartigen Tumore korrekt

False-positive-rate (FPR): FP/FP+TN oder FP/N
Spannungen zwischen Precision und recall

Darum gibt es F1- Score als ein kombiniertes Maß für Precision und Recall
F1 = 2 * (precision*recall) / precision +recall)

- zeigt die precision recall Kurve

ROC curve: Receiver operating characteristics

- y-Achse: True positive rate (oder recall), x-Achse: False positive Rate
- AUC ist ein Maß für Klassifikationsperformance, welches unabhängig von der Entscheidungsregel ist
- Perfekter Klassifikator = 1, no skill = 0

Naive-Bayes Klassifikator

- Spamfilter Beispiel

Woche 8: Unsupervidiertes Lernen

- Meist für explorative Datenanlyse verwendet
- Wir verwenden Prädikotren aber keine Responsevariabel
- Ziel a) Erkennen von Zusammenhängen zwischen Variablen und Observationen/Individuen b) erkenne von Muster auf unsupervidierte Art
- Geeignet für p > n Szenarien, mehr Messungen als Probanden
- Dimensionsreduktion für hoch-dimensionale Daten
- Problem: performance ist schwer zu beurteilen, da es keine Validierungsmnethoden gibt, subjektive Interpretation

Principal component analysis

- Reduktion der Dimensionen
- Finde Hauptkomponenten, die die größte Varianz der Daten aufzeigen
- Plotte die Hauptkomponenten orthogonal in einem neuen Raum

PCA: Schritt1

- Zentriere die Daten am Mittelwert x = x –mean(x)

PCA: Schritt 2

- Finde Richtung der größten Varianz = PC1

PCA: Schritt 3

- Finde Richtung der zweitgrößten Varianz, die orthogonal (kreuzend) zu PC1 ist = PC2

PCA Schritt 4

- Transformiere Daten in neues Koordinatensystem mit PC1 und PC2

Scree-Plot: zeigt Anteil der von jeder PC erklärten Varianz

Hierarchisches Clustering:

- Bottum up: agglomerativ, starten von unten nach oben, jedes Sample startet in eigenen cluster, ähnliche werden gemerged
- Top down: alle starten in einem Cluster und werden recursiv aufgespalten in mehrere cluster
- Basieren auf einem **Dissimilaritätsmaß,** bestehend aus einer Distanzmetrik und einem Linkage Kriterium

Linkage Kriterium:

- Single oder complete
- Bestimmt die Distanz zwischen Observationssets als Funktion der paarweisen Distanz zwischen Observationen

Beispiel bottum up & complete linkage:

- **Suche das Paar a und b mit der geringsten Distanz**
- A und b werden mit Knotenpunkt u verbunden
- Nun werden alle Distanzen zwischen dem cluster zwischen (a,b) und allen anderen Klassen
- Maximale Distanz soll zwischen jedem Elemtnt und dem cluster gewählt werden
- Unterschied: single linkage sucht die kleinste Distanz

K-means cluster

- Ordne Observationen einer festen Anzahl an K Clustern zu, wobei die Anzahl k voher gewählt wird
- 1) ordne Observationen dem nähesten Cluster zu
- 2) berechne Gravitätszentrum jedes Cluster

Probleme:

- Lösung hängt vom Start der Cluster ab
- Mehrmals mit unterschiedlichen k durchführen
- Überprüfe mit silhouette plots
- Ausreißer werden mit einberechnet

Alternative DBSCAN:

- Identifiziere Cluster anhand der Dichte der Datenpunkten
- 1) zähle für *jeden* Datenpunkt die Anzahl an Punkten die am nächsten liegen (in Distanzmaß gemessen)
- 2) definiere core punkte, die mindestens 4 nahe Datenpunkte haben (4 ist frei gewählt)
- 3) wähle zufällig 1 corepoint und ordne diesem einen Cluster zu
- 4) core points die in der Nähe sind, werden dem Cluster zugewiesen
- 5) nicht-core points werden nur zur erwieterung nahe geliegender cluster genutzt
- Alle übrigen nicht-core punkte werden als ausreißer behandelt

Vorteil:

- Muss nicht vorher clusteranzahl definieren
- Kann beliebige Formen annehmen

Nachteil:

- Datensätze mit unterschiedlichen Dichten können zu Problemen führen
- Nachbarschaftslänge und mindestanzahl der nachbarpunkte der core points müssen definiert werden

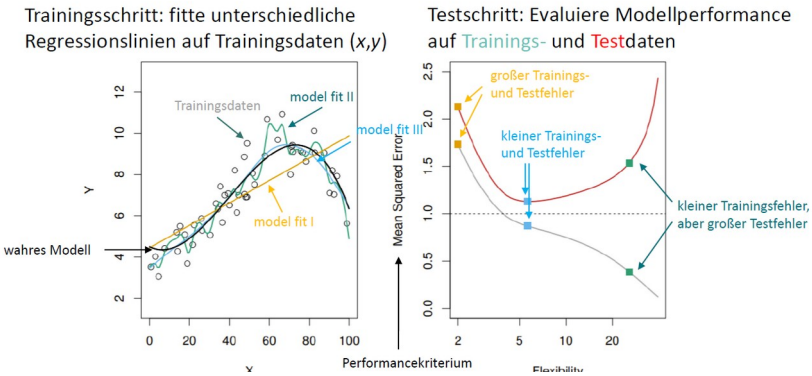
Woche 9: Overfitting, Regularisierung und Entschiedungsbäume

Grundproblem: Wenn die Anzahl an Trainingssamples **kleiner** ist als die Anzahl der Prädiktorvariablen, ist das Gleichungssystem nicht lösbar

- Dies ist oft im omics Bereich der Fall
- Gefahr des overfittings ist sehr groß

Overfitting

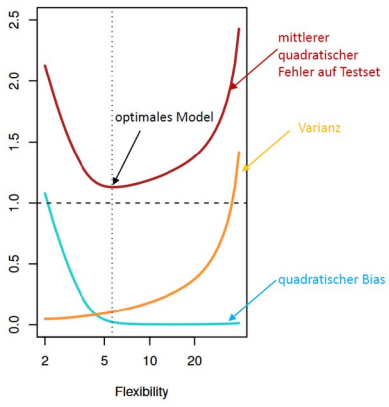
Im Trainingsschritt wird ein Model auf Trainigsdaten gefittet. Danach wird im Testschritt die Modellperformance mit Trianings- und Testdaten evaluiert



Modelfit II eher komplex/sehr flexibel und beschreibt, Trainingsdaten sehr exakt (sensitiv bzgl. Rauschen/Noise in Trainingsdaten), aber beschreibt Testdaten sehr schlecht → Overfitting = schlechte Modellperformance auf
neuen Daten + schlechte Generalisierbarkeit

Modelfit I eher einfach/weniger flexibel, beschreibt Trainingsdaten nicht gut und beschreibt Testdaten ebenfalls schlecht → Underfitting = schlechte Modellperformance

Bias- variance Trade off



Variance: Fehler aufgrund eines zu sensitven Modells, welches sensitiv bezüglich kleiner Fluktuationen ist (Overfitting)

Bias: „Verzerrung“ aufgrund eines zu einfachen Modells (Underfitting)

Optimale Model: Trade off zwischen Bias und Variance

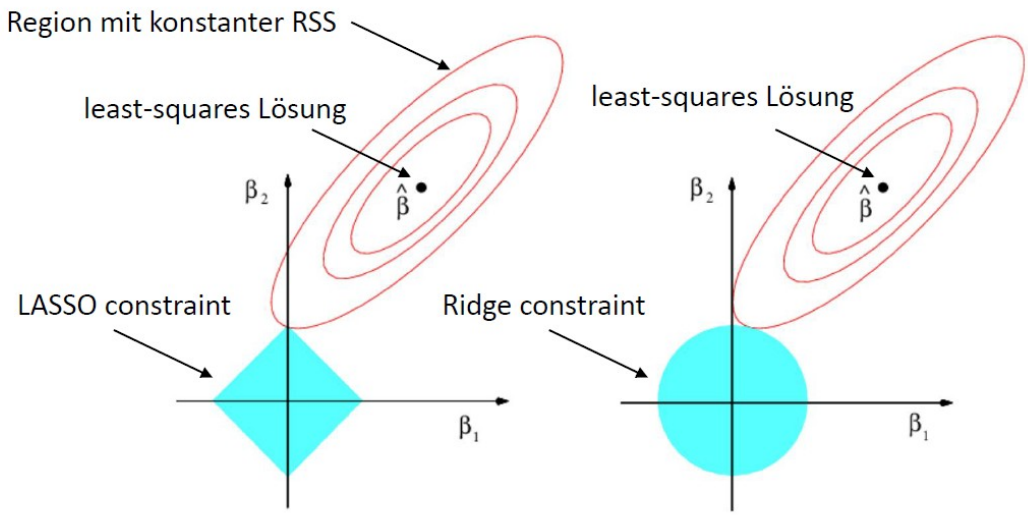
Lösung 1: Immer Trainingsdaten benutzen
Lösung 2: Variablenselektion durch automatische Variablenselektion
Lösung 3: reduziere Variablenraum mittels Dimensionsredution z.B. PCA
Lösung 4: Regularisierungsmethoden (z.B. LASSO)

Regularisierungsmethoden

Prinzip: schrumpfe/regularisiere die Modellkoeffizienten (um Overfitting zu vermeiden), indem eine spezifische Bedingung („constraint“) erfüllt werden muss

Durch: LASSO-Regularisierungsterm I1-Norm oder lamda oder Ridge I2-Norm Regularisierungsterm

- Lamda wird optimiert (>0) durch interne Crossvalidierung



- LASSO und Ridge schrumpfen die absoluten Werte der Modellkoeffizienten Richtung Null und vermeiden damit Overfitting
- Vorteil von LASSO: einige Variablen werden auf 0 gedrückt und damit aus dem Modell gekickt
- Lamda min ist das optimale min mit bestem Optimierungsmaß, gibt eine Anzahl der Variablen an, Lamda
- Lamda 1sd eine Standardabweichung

Entscheidungsbäume:

- Root Knoten: Ausgangsfrage
- Interne Knoten: Branch geht rein und branch geht raus
- Leaf Knoten: branch geht rein, keiner geht raus

Welche Prädiktorvariable wird root?

- Alle Variablen werden evaluiert hinsichtlich des Vorkommens in der binären Responsevariable
- Leaf notes können unrein sein, es kommt nicht 100% einer Kategorie in einem Knoten vor
- Dies Unreinheit/impurity kann gemessen werden als Gini-Index

Formel GINI-Index

Ablauf des Entschiedungsbaumes:

- Jede Prädikorvariable bekommt einen GINI-Index

- Der Knoten, der den kleinsten GINI-Index hat i.e., die kleinste Impurity, wird als root node gesetzt
- In den neuen Branchen werden die restlichen Prädiktoren evaluiert und nach GINI Index aufgelistet
- Wird der GINI-Index nicht kleiner in einem Branch, dann wird an dieser Stelle der Branch als leaf Knoten übriggelassen
- Kann nominal oder kontinuierlich sein, bei letzteren werden die Mittelwerte zweier Samples berechnet und der GINI Index berechnet (Treshold mit dem niedrigsten GINI wird als root genommen
- Meist schlechte Prädiktionsperofrmance, daher wird oft ein random forest gebaut

Random forest

1. Bootstrappe die Trainingsdaten, d.h. teile sie in kleine randomisierte Sub Sets aus
2. Baue einzelne Bäume auf Trainingsdaten
3. Wiederhole alles
4. Responsewerte der neuen Testdaten werden dann mittels Mehrheitsbeschlusses über alle Bäume hinweg vorhergesagt

Altklausurfragen

Wie kann man eine metrische Variable darstellen?

- Bar Chart, Boxplot, Scatter plot

1. Definition folgender Variablentypen: Binär, Kategorial, Metrisch, ordinal

2. Ordnen sie folgenden Variablen Variabelntypen und -subtypen zu

    a. Gewicht in kg -> metrisch
    b. Besuche der Notaufnahme -> diskret
    c. Fieber ja/nein -> binär
    d. Symptome (Husten, Fieber, Atemnot..) -> kategorial
    e. Wie häufig Symptome (Immer, meistens, oft, selten, nie) -> Ordinal

3. Jew. 2 Maßzahlen für Normale Verteilung (Mittelwert & SD) und schiefe Verteilung (Median und IQR)
4. Was ist git/wozu kann man git nutzen
    a. Versionsverwaltung
    b.
5. Wie heißt der Befehl in git, um die gestackten Änderungen vor dem push zu sammeln?
    a. Git pull
    b. Git push
    c. Git commit (x)
6. Nennen sie zwei Softwares, die sie bei der Datenverarbeitung unterstützen

Woche 2

1. Risiko, Riskratio, Odds und Odds Ratio berechnen

Woche 3

Sensitivität und Spezifität berechnen (2P)

Woche 4

1. Sortierung der Visualisierungsmethoden nach Anzahl der darzustellenden Variablen

- Gauge Chart, Scatter plot, Bubble plot,

Woche 5

1. Wie können multivariate Daten dargestellt werden
    a. Glyph (x)
    b. Figur
    c. Symbol
2. Wie können qualitative Daten graphisch dargestellt werden?
    a. Mosaik Matrix
    b. Scatterplot
    c.

Woche 6

1. 4x5 tibble -> Filter und select kennen und Zeilen & Spalten bleiben übrig
   Select nach Spaltenname oder einzelnem ‚R'

Woche 7 Lineare Regression

Woche 8 Überwachte Methoden

1. Nennen Sie vier Methoden des überwachten Lernens: z.B. k-Neares Neighbor, LDA, QDA, SVM

Woche 9 Unüberwachte Methoden

1. Was trifft zu
   a. PCA und LDA zählen zu den unüberwachten Methoden
   b. KNN und LDA zählen zu den unüberwachten Methoden
   c. Hierarchisches Clustering und PCA zählen zu den unüberwachten Methoden (x)
   d. K-means und PCA zählen zu den unüberwachten Methoden (x)
2. Assoziationsanalyse (Lückentext)
   a. Support
   b. Konfidenz
   c. Die Rechenleistung steigt mit mehr Variablen
   d. Der Apriori-Algorithmus eignet sich für kleine Mengen
   e. Support und Konfidenz sollen möglichst hoch sein
3. Was trifft zu
   a. Unüberwachte Methoden differenzieren nicht zwischen abhängiger und unabhängiger Variable (x)
   b.