[1 Introduction 9](#_Toc109376347)

[1.1 What are x and y called in models? 9](#_Toc109376348)

[1.2 What are the two most basic types of models? 9](#_Toc109376349)

[1.3 How can regression problems be transformed into classification problems? 9](#_Toc109376350)

[1.4 Describe the two basic motives why you would want to build a model. 9](#_Toc109376351)

[1.5 Describe the workflow that is often done in statistical machine learning. 9](#_Toc109376352)

[1.6 What is one basic trade-off when choosing models? How do some models score on this trade-off? 9](#_Toc109376353)

[1.8 What does X look like in formula form? What is the relationship between X and Y? What do the terms mean? 10](#_Toc109376354)

[1.9 What is the main difference between non-parametric and parametric methods? Describe the advantages and disadvantages. 10](#_Toc109376355)

[1.10 You have fitted a model. How would you assess model quality in practice? 10](#_Toc109376356)

[1.11 What is the formula for the training MSE? What is the formula for the test MSE? 10](#_Toc109376357)

[1.12 What does the expected test MSE decompose as? 11](#_Toc109376358)

[1.13 What parts are reducible? What parts are irreducible? 11](#_Toc109376359)

[1.14 Describe the bias-variance trade-off. What kind of models are high-bias and high-variance? What is the danger here? 11](#_Toc109376360)

[1.15 Illustrate a model with high, medium and low bias or low, medium and high variance. 11](#_Toc109376361)

[1.16 What performance measure would you use for classification? 11](#_Toc109376362)

[1.17 What is the formula for the error rate? 11](#_Toc109376363)

[1.18 Describe the Bayes classifier. What is its function? 11](#_Toc109376364)

[1.19 What is the formula of the Bayes error rate? 11](#_Toc109376365)

[1.20 What is the formula for the k-nearest neighbor classifier? 12](#_Toc109376366)

[1.21 Describe the difference between supervised and unsupervised learning? 12](#_Toc109376367)

[1.22 What are other types of learning? 12](#_Toc109376368)

[2 Linear Regression 13](#_Toc109376369)

[2.1 What are the two parameters/coefficients of simple linear regression? 13](#_Toc109376370)

[2.2 What is the formula for simple linear regression? 13](#_Toc109376371)

[2.3 What is the formula for a residual? 13](#_Toc109376372)

[2.4 What is the formula for the residual sum of squares? 13](#_Toc109376373)

[2.5 What is the analytical computation for ordinary least squares parameters? 13](#_Toc109376374)

[2.6 What is the formula for the residual standard error? What is its function? How does it change in size and? 13](#_Toc109376375)

[2.7 What is the formula for the R2? What are the components of the formula? What does it mean? 13](#_Toc109376376)

[2.8 For simple linear regression, what does the R2 equal? 13](#_Toc109376377)

[2.9 What is the formula for the standard error for the intercept and slope? 13](#_Toc109376378)

[2.10 What are the 95%-confidence intervals approximately equal? 14](#_Toc109376379)

[2.11 What is the correct interpretation of confidence intervals? 14](#_Toc109376380)

[2.12 What are the general formulations for null hypothesis and alternative hypothesis? 14](#_Toc109376381)

[2.13 How would you determine whether the Null hypothesis can be rejected? 14](#_Toc109376382)

[2.14 How do you get the p-value from a t-value? 14](#_Toc109376383)

[2.15 Give one definition of degrees of freedom? 14](#_Toc109376384)

[2.16 Example: You have five observations and calculate the mean. How many degrees of freedom does the mean have? 15](#_Toc109376385)

[2.17 Describe model degrees of freedom. 15](#_Toc109376386)

[2.18 How many model degrees of freedom do linear regression and kNN regression have? 15](#_Toc109376387)

[2.19 What are residual degrees of freedom? What is its relationship with model degrees of freedom? 15](#_Toc109376388)

[2.20 What is the frequentist approach to probability? 15](#_Toc109376389)

[2.21 What is the Bayesian approach to probability? 15](#_Toc109376390)

[2.22 What is the formula of Bayes’ Theorem? 15](#_Toc109376391)

[2.23 What is the technical difference when predicting using frequentist approaches or Bayesian approaches? 15](#_Toc109376392)

[2.24 Correlation is not causation. 15](#_Toc109376393)

[2.25 What is the formula for multiple linear egression? 16](#_Toc109376394)

[2.26 What is the formula for the residual sum of squares in multiple regression? 16](#_Toc109376395)

[2.27 What is the residual sum of squares defined as for multiple linear regression? 16](#_Toc109376396)

[2.28 What is the null and alternative hypothesis for multiple linear regression? 16](#_Toc109376397)

[2.29 What is the formula for the F-statistic? 16](#_Toc109376398)

[2.30 When do the numerator and denominator for the F-statistic become sigma squared? 16](#_Toc109376399)

[2.31 How can we test for significance? 16](#_Toc109376400)

[2.32 Shortly describe forward selection, backward and mixed selection. 16](#_Toc109376401)

[2.33 What are dummy variables? What do they change? 17](#_Toc109376402)

[2.34 Describe interaction terms. 17](#_Toc109376403)

[2.35 What is the formula for polynomial regression? 17](#_Toc109376404)

[2.36 What are different examples for generalized linear models? 17](#_Toc109376405)

[2.37 What is it called when you have a cone shape in your residual plot? Simple fix? 17](#_Toc109376406)

[2.38 What is an outlier? What is its effect on the regression fit? And on the residual sum of squares? 17](#_Toc109376407)

[2.39 How can you identify outliers? What to do? 17](#_Toc109376408)

[2.40 What are high leverage points? 17](#_Toc109376409)

[2.41 What is a criterion for high leverage? 17](#_Toc109376410)

[2.42 What is collinearity? What to do? 17](#_Toc109376411)

[3 Classification 18](#_Toc109376412)

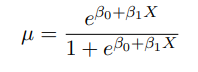
[3.1 What are problems with linear regression when you have categorical features? 18](#_Toc109376413)

[3.2 What is the formula for simple logistic regression? 18](#_Toc109376414)

[3.3 How does logistic regression work? 18](#_Toc109376415)

[3.4 What is the formula difference between the logistic function and the logit function? What do they look like? 18](#_Toc109376416)

[3.5 What do they odds and log odds ratio mean in the case of logistic regression? What values do they take? 18](#_Toc109376417)

[3.6 What does the mean function look like for logistic regression?  19](#_Toc109376418)

[3.7 INSERT HERE A QUESTION ABOUT LIKELIHOOD AND MAXIMUM LIKELIHOOD THAT I DON’T UNDERSTAND YET 19](#_Toc109376419)

[3.8 How do you estimate a classification out of probabilities from the logistic regression? 19](#_Toc109376420)

[3.9 INSERT HERE A QUESTION ABOUT MULTINOMIAL LOGISTIC REGRESSION THAT I DON’T UNDERSTAND YET 19](#_Toc109376421)

[3.11 Describe distriminative learning approaches. 19](#_Toc109376422)

[3.12 Describe generative learning approaches. 19](#_Toc109376423)

[3.13 What is the formula for the gaussian distribution? 20](#_Toc109376424)

[3.14 What is the idea behind linear discriminant analysis`? 20](#_Toc109376425)

[3.15 How are the parameters for linear discriminant analysis estimated via the maximum likelihood principle? 20](#_Toc109376426)

[3.16 How does quadratic discriminant analysis differ from linear discriminant analysis? 20](#_Toc109376427)

[3.17 What is the underlying assumptions for the Naïve Bayes classifier? 20](#_Toc109376428)

[3.18 What is the formula for the naïve bayes classifier? 21](#_Toc109376429)

[3.19 What are the estimates for pi and theta for the Naïve bayes classifier? 21](#_Toc109376430)

[3.20 What is a key advantage of naïve bayes? 21](#_Toc109376431)

[3.21 What are some properties of Naïve Bayes? 21](#_Toc109376432)

[3.22 What is the formula for a Bayesian network based on a directed acyclic graph G? 21](#_Toc109376433)

[3.23 Describe a Tree Augmented Naive Bayes. 21](#_Toc109376434)

[3.24 What are the differences between logistic regression and linear discriminant analysis? 21](#_Toc109376435)

[3.25 When is Naïve Bayes good 22](#_Toc109376436)

[3.26 Is discriminative learning always better than generative learning? 22](#_Toc109376437)

[3.27 What is the difference between quadratic discriminant analysis and tree augmented bayes and LDA and Naïve Bayes? 22](#_Toc109376438)

[4 Performance evaluation 23](#_Toc109376439)

[4.1 What are the goals of performance evaluation? 23](#_Toc109376440)

[4.2 What are challenges? 23](#_Toc109376441)

[4.3 Name three measures used for evaluating regression tasks and their formulas. 23](#_Toc109376442)

[4.4 Which measures penalize larger deviations from ground truth more harshly? 23](#_Toc109376443)

[4.5 What measure is sufficient as a loss function for optimization? 23](#_Toc109376444)

[4.6 What are the components of a confusion matrix? 24](#_Toc109376445)

[4.7 What are the formulas for sensitivity, specificity and precision? 24](#_Toc109376446)

[4.8 What is the formula for the F1 score? What are its properties? 24](#_Toc109376447)

[4.9 What is the formula for Matthews’ correlation coefficient? What are its properties? 24](#_Toc109376448)

[4.10 How does a receiver operating characteristic (ROC) work? What does it look like? What are the labels on the x and y axis? 25](#_Toc109376449)

[4.11 How does a precision-recall curve work? What does it look like? What are the labels on the x and y axis? 25](#_Toc109376450)

[4.12 How can you transform the confusion matrix of a multi classification problem into many binary classification problems? 25](#_Toc109376451)

[4.13 Describe the holdout approach for data set splitting into training and test set. What are its problems? 25](#_Toc109376452)

[4.14 Describe the process off K-fold cross validation. 25](#_Toc109376453)

[4.15 Describe leave-one-out cross validation. What is the analytical solution for linear/polynomial regression? 26](#_Toc109376454)

[4.16 What should you consider when choosing k? 26](#_Toc109376455)

[4.17 What is pooling in regards to performance measures? 26](#_Toc109376456)

[4.18 What is pooling in regards to the original predictions? 26](#_Toc109376457)

[4.19 Describe repeated holdout. What is the difference to k-fold CV? 26](#_Toc109376458)

[4.20 What are other repetition strategies and how do they work? 26](#_Toc109376459)

[4.21 You have obtained the average errors for your evaluation strategy. Which model should you choose? 27](#_Toc109376460)

[4.22 What is the problem with performing simple hold-out evaluation for tuning a hyperparameter hat q. 27](#_Toc109376461)

[4.23 Describe (with illustration) the process of nested cross validation. 27](#_Toc109376462)

[4.24 What would be the ideal process or quantifying uncertainty of an estimate? Why is that not possible? 27](#_Toc109376463)

[4.25 What is the idea and process of bootstrapping? 27](#_Toc109376464)

[5 Regularization 28](#_Toc109376465)

[5.1 What are the advantages of regularizing linear models? 28](#_Toc109376466)

[5.2 Name three methods for regularization and briefly describe them. 28](#_Toc109376467)

[5.3 Name three methods one can use to choose between models of different complexity (number of features). 28](#_Toc109376468)

[5.4 Describe the process of forward stepwise selection. 28](#_Toc109376469)

[5.5 What are advantages and disadvantages of forward stepwise selection? 28](#_Toc109376470)

[5.6 What is the complexity of forward stepwise selection? 28](#_Toc109376471)

[5.7 What is the process of backward stepwise selection? 28](#_Toc109376472)

[5.8 What are the complexity and sample requirements of backward stepwise selection? 29](#_Toc109376473)

[5.9 What is the basic formula for shrinkage methods? 29](#_Toc109376474)

[5.10 Why do shrinkage methods work? 29](#_Toc109376475)

[5.11 Name some variants for shrinkage methods 29](#_Toc109376476)

[5.12 What formula is used for g in ridge regression? What do the parameters mean? 29](#_Toc109376477)

[5.13 What are the main disadvantages of ridge regression? 29](#_Toc109376478)

[5.14 Draw an example for the course of variance, squared bias and test MSE on real data for lambda and l2-penalty. 29](#_Toc109376479)

[5.15 Describe the formulafor lasso. What do te parameters mean? 30](#_Toc109376480)

[5.16 Conceptually, what is the main difference between Lasso and Ridge Regression and Subset selection? 30](#_Toc109376481)

[5.17 What are the geometries of the l1 and the l2 penalties? 30](#_Toc109376482)

[5.18 What does it look like when you plot lambda and R2 on training data by mean squared error for Ridge regression and lasso? 31](#_Toc109376483)

[5.19 What does it look like when you plot lambda and R2 on training data by mean squared error for Ridge regression and lasso and some coefficients are exactly zero? 31](#_Toc109376484)

[5.20 What is the formula for the penalty term for Elastic net? 31](#_Toc109376485)

[5.21 Describe the process of parameter tuning. 31](#_Toc109376486)

[6 Decision Trees 33](#_Toc109376487)

[6.1 Describe the structure and parts of decision trees. 33](#_Toc109376488)

[6.2 What are the two kinds of decision trees? What do the leaves store? 33](#_Toc109376489)

[6.3 What are the advantages and disadvantages of decision trees? 33](#_Toc109376490)

[6.4 How can one evaluate the quality off a region? 33](#_Toc109376491)

[6.5 Name some impurity measures for decision trees. What are their formulas? 34](#_Toc109376492)

[6.6 What do the different impurity measures have a course like depending on the probability of class? 34](#_Toc109376493)

[6.7 What approaches are there to control tree complexity? 34](#_Toc109376494)

[6.8 When fitting a regression tree instead of a classification tree, how is it done? 34](#_Toc109376495)

[6.9 How is the quality of tree T defined? 34](#_Toc109376496)

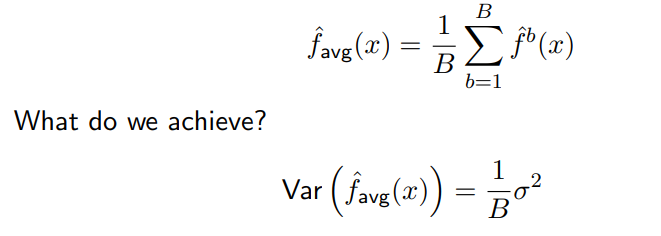
[6.10 What are characteristics of this quality measure? 35](#_Toc109376497)

[7 Ensemble Learning 36](#_Toc109376498)

[7.1 What are the two kinds of ensemble learnings? 36](#_Toc109376499)

[7.2 What are the two types of ensemble methods? 36](#_Toc109376500)

[7.3 How does the average variance of f models trained on B data sets change in respect to the original variance? 36](#_Toc109376501)

[ 36](#_Toc109376502)

[7.4 How does bagging work? What kind of ensemble does it produce? 36](#_Toc109376503)

[7.5 What are the limits of bagging? 36](#_Toc109376504)

[7.6 What is the variance of bagging? What does increasing B do? 36](#_Toc109376505)

[7.7 Explain how to out of bag error. 36](#_Toc109376506)

[7.8 What is the idea of random forests? 37](#_Toc109376507)

[7.9 What is the aim of random forests? 37](#_Toc109376508)

[7.10 Describe the process to learn a random forest. 37](#_Toc109376509)

[7.11 What are the two formulas that a sample hat y\_i is predicted by by regression and classification? 37](#_Toc109376510)

[7.12 What are the two most important hyperparameters and their default values? 37](#_Toc109376511)

[7.13 When should you stop splitting for regression trees? 37](#_Toc109376512)

[7.14 Is boosting building a sequential or a parallel ensemble? 37](#_Toc109376513)

[7.15 Name some characteristics of boosting. 37](#_Toc109376514)

[7.16 Describe the process of adaptive boosting. 37](#_Toc109376515)

[7.17 What is the formula for the weighted majority vote in Boosting and what are its parameters? 38](#_Toc109376516)

[7.18 How does AdaBoost for regression work? 38](#_Toc109376517)

[7.19 What function do we minimize in gradient bosting? 38](#_Toc109376518)

[7.20 What differences are there between gradient boosting and XGBoost? 38](#_Toc109376519)

[7.21 How can one interpret variable importance in a random forest? 38](#_Toc109376520)

[7.22 What is one word of caution when interpreting variable importance? 38](#_Toc109376521)

[7.23 How could one visualize the relative importance between two features? 38](#_Toc109376522)

[8 Support Vector Machines 39](#_Toc109376523)

[8.1 How do support vector machines separate two classes? How do they achieve nonlinearity? 39](#_Toc109376524)

[8.2 What two special cases of SVMs did we have in lecture? 39](#_Toc109376525)

[8.3 What is a hyperplane? 39](#_Toc109376526)

[8.4 How does one come to a decision regarding y\_i with a hyperplane? 39](#_Toc109376527)

[8.5 How do you find an optimal separating hyperplane? 39](#_Toc109376528)

[8.6 How can you solve for the maximal separating hyperplane? 39](#_Toc109376529)

[8.7 What are problems of the maximal margin classifier? 39](#_Toc109376530)

[8.8 What is a soft margin in a support vector machine? 39](#_Toc109376531)

[8.9 Describe the optimization problem for the support vector classifier. 40](#_Toc109376532)

[8.10 What are slack variables and what are the different cases for slack variables? 40](#_Toc109376533)

[8.11 What does it mean if C has a high/low value? How is C chosen? 40](#_Toc109376534)

[8.12 What is one solution to fit SVCs on non-linear data without using kernels? 40](#_Toc109376535)

[8.13 How does a general Support Vector Machine work? 40](#_Toc109376536)

[8.14 Name some kernels and reproduce their formulas. 40](#_Toc109376537)

[8.15 What is the advantage of using kernels vs. an enlarged feature space? 41](#_Toc109376538)

[8.16 What should you do if you have more than two classes? 41](#_Toc109376539)

[9 Clustering 42](#_Toc109376540)

[9.1 Is Clustering a supervised or unsupervised learning approach? 42](#_Toc109376541)

[9.2 What are the goals of cluster analysis? 42](#_Toc109376542)

[9.3 What is dissimilarity (with formula)? 42](#_Toc109376543)

[9.4 What are common dissimilarity functions? 42](#_Toc109376544)

[9.5 What are the ideas of k-means? 42](#_Toc109376545)

[9.6 What are the minimization problem formulas for k-means? 42](#_Toc109376546)

[9.7 Why is finding the globally optimal solution for k-means hard? 42](#_Toc109376547)

[9.8 What is the algorithm for finding a locally optimal solution? 42](#_Toc109376548)

[9.9 What is the k-means partitioning called? 42](#_Toc109376549)

[9.10 What is a common problem for k-means classification? 42](#_Toc109376550)

[9.11 How can one choose the optimal K? 43](#_Toc109376551)

[9.12 What is the process for a Gaussian mixture model? 43](#_Toc109376552)

[9.13 Explain K-medoids. 43](#_Toc109376553)

[9.14 What is the idea of hierarchical clustering? 43](#_Toc109376554)

[9.15 What is the process of agglomerative clustering? 43](#_Toc109376555)

[9.16 What is the formula and what are the characteristics of single linkage? 43](#_Toc109376556)

[9.17 What is the formula and what are the characteristics of complete linkage? 44](#_Toc109376557)

[9.18 What is the formula and what are the characteristics of average linkage? 44](#_Toc109376558)

[9.19 What is a particular use case for agglomerative clustering? 44](#_Toc109376559)

[9.20 What is the process for reconstructing a phylogenetic tree? 44](#_Toc109376560)

[9.21 What is the process of divisive clustering? 44](#_Toc109376561)

[9.22 Why is divisive clustering rarely used in practice? 44](#_Toc109376562)

[10 Dimension reduction 45](#_Toc109376563)

[10.1 What is the goal of dimensionality reduction? 45](#_Toc109376564)

[10.2 What is the motivation for dimension reduction? 45](#_Toc109376565)

[10.3 Name dimension reduction methods 45](#_Toc109376566)

[10.4 What is the general idea of principal component analysis? 45](#_Toc109376567)

[10.5 What preprocessing should be done to the data before PCA? 45](#_Toc109376568)

[10.6 What is the formula of the first principal components? What are the factors called? 45](#_Toc109376569)

[10.7 What value constraint do the loadings have? 46](#_Toc109376570)

[10.8 How do we solve for the first principal component? What formula do we maximize? 46](#_Toc109376571)

[10.9 What conditions does the second principle component need to fulfil? 46](#_Toc109376572)

[10.10 What is the formula for the proportion of variance explained? 46](#_Toc109376573)

[10.11 What does a Scree plot show? Draw an example. 46](#_Toc109376574)

[10.12 What is the main difference between PCA and t-SNE? 46](#_Toc109376575)

[10.13 What are other approaches that preserve small pairwise distances? 47](#_Toc109376576)

[10.14 What are the parameters of t-SNE and what is the process for its optimization? 47](#_Toc109376577)

[10.15 How should you choose P? 47](#_Toc109376578)

[10.16 What is perplexity? 47](#_Toc109376579)

[10.17 What is Q? 48](#_Toc109376580)

[10.18 What is the formula of the Kullback-Leibler divergence? What does it measure? 48](#_Toc109376581)

[10.19 How is the Kullback-Leibler divergence asymmetric? What does that mean for different values for p and q? 48](#_Toc109376582)

# Introduction

## What are x and y called in models?

* x: feature, variable, input, independent variable, ...
* y: response, target, dependent variable, …

## What are the two most basic types of models?

Regression and Classification

## How can regression problems be transformed into classification problems?

Discretize the response

## Describe the two basic motives why you would want to build a model.

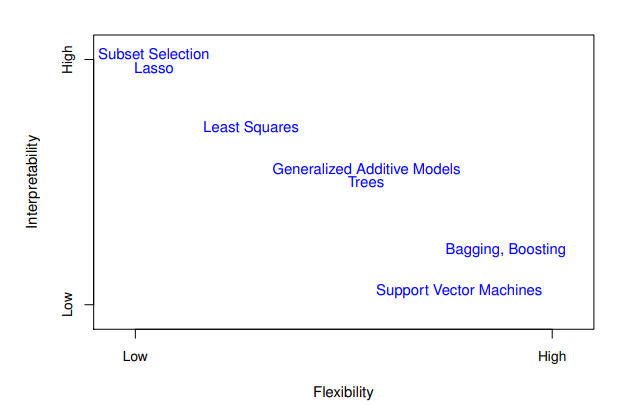
* Prediction:
  + Forecasting the response for new features values
  + The machine learning viewpoint
* Inference
  + Creating a model of the data-generating process
    - Which features are associated with the response?
    - How do important features influence the response?
  + The statistics viewpoint

## Describe the workflow that is often done in statistical machine learning.

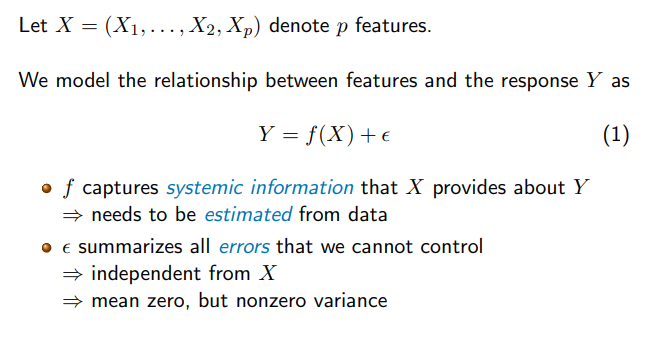
* Learn a model that optimizes predictive performance
* Perform inference on the learned model

## What is one basic trade-off when choosing models? How do some models score on this trade-off?

Interpretability and flexibility



## What does X look like in formula form? What is the relationship between X and Y? What do the terms mean?



## What is the main difference between non-parametric and parametric methods? Describe the advantages and disadvantages.

Parametric methods:

* make an assumption about the functional form of f.
* We use the training data to estimate the model parameters 🡪 we obtain a trained model hat f
  + Advantage: Estimating parameters is easier than estimating functions
  + Disadvantage: model assumptions may be very wrong

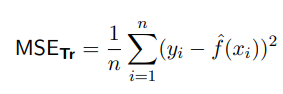
Non-Parametric models:

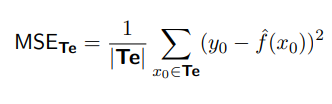
* Seek estimate of f that gets close to the training data points
* Some constraints on complexity of the fit
  + Advantage: No restrictions w.r.t. form of f
  + Disadvantage: A lot of data needed to fit f

## You have fitted a model. How would you assess model quality in practice?

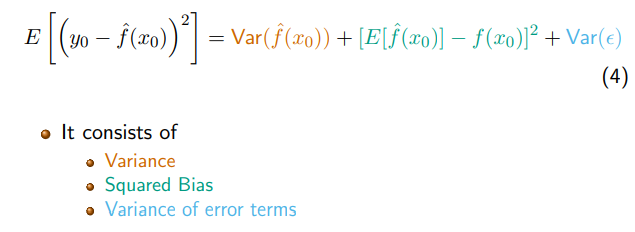
Compute the Mean squared error on Test data.

## What is the formula for the training MSE? What is the formula for the test MSE?





## What does the expected test MSE decompose as?



## What parts are reducible? What parts are irreducible?

* Variance and squared bias are reducible
* The variance of eta is irreducible (missing features, inherent noise, other factors)

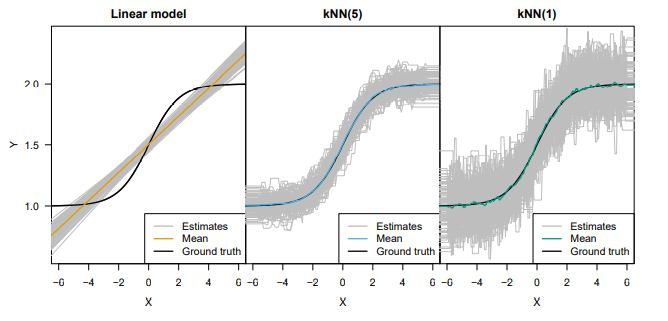
## Describe the bias-variance trade-off. What kind of models are high-bias and high-variance? What is the danger here?

The bias-variance tradeoff that a flexible model has low bias but by definition high variance, whereas an inflexible model has low variance but high bias.

The danger is over- bzw. Underfitting.

Examples: Flexible: knn(1), Inflexible: Linear regression

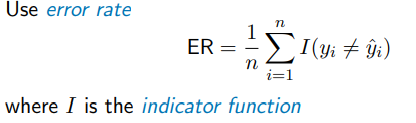
## Illustrate a model with high, medium and low bias or low, medium and high variance.



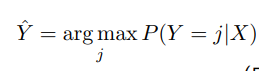
## What performance measure would you use for classification?

Error rate.

## What is the formula for the error rate?



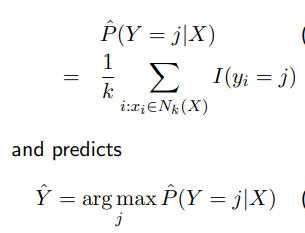
## Describe the Bayes classifier. What is its function?



## What is the formula of the Bayes error rate?



## What is the formula for the k-nearest neighbor classifier?



## Describe the difference between supervised and unsupervised learning?

* Supervised: For every feature x\_i we have a response y\_i
* Unsupervised: Missing response data. 🡪 Clustering, dimensionality reduction

## What are other types of learning?

* Semi-supervised: Some but not all y\_i available
* Reinforcementt learning: Maximize a reward function

# Linear Regression

## What are the two parameters/coefficients of simple linear regression?

Intercept and slope

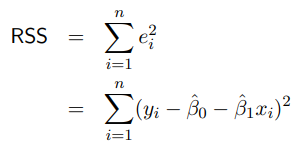
## What is the formula for simple linear regression?



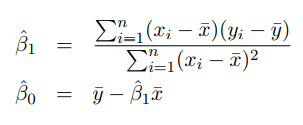
## What is the formula for a residual?



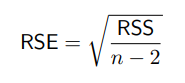
## What is the formula for the residual sum of squares?



## What is the analytical computation for ordinary least squares parameters?

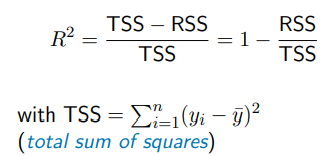


## What is the formula for the residual standard error? What is its function? How does it change in size and range?



* Estimates the standard deviation of eta
* Value range depends on range of Y (small value is good)

## What is the formula for the R2? What are the components of the formula? What does it mean?

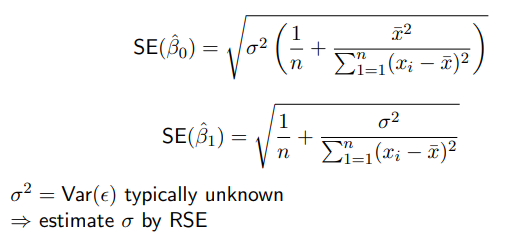


* Fraction of variance explained (values between 0 and 1)
* High value is good

## For simple linear regression, what does R2 equal?



## What is the formula for the standard error for the intercept and slope?



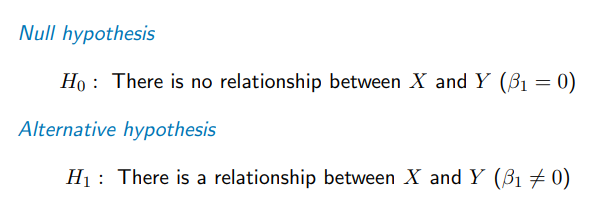
## What are the 95%-confidence intervals approximately equal?

Hat intercept/slope +- 2\*SE(intercept/slope)

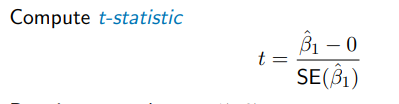
## What is the correct interpretation of confidence intervals?

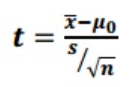
The confidence interval has a 95% chance of actually containing the true parameter.

## What are the general formulations for null hypothesis and alternative hypothesis?



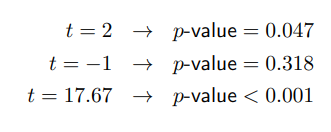
## How would you determine whether the Null hypothesis can be rejected?



Generally: 

## How do you get the p-value from a t-value?

P-value is defined as observing any value >= |t| 🡪 small p-value: null hypothesis is rejected

For reference of magnitudes: 

## Give one definition of degrees of freedom?

The degrees of freedom of a system is the number of parameters of the system that may vary independently.

## Example: You have five observations and calculate the mean. How many degrees of freedom does the mean have?

5-1. This is because you have to choose the last value of a hypothetical second distribution so that it fits the mean

## Describe model degrees of freedom.

* Measures the flexibility in choosing the model
* Allows comparing model complexity across very different types of models
* Small values mean less danger of overfitting

## How many model degrees of freedom do linear regression and kNN regression have?

* Linear regression: p
* kNN regression: n/k

## What are residual degrees of freedom? What is its relationship with model degrees of freedom?

* Degree of flexibility in choosing an equivalent data set that satisfies model assumptions.
* Larger values are good
* Residual df = n – model df

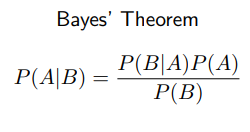
## What is the frequentist approach to probability?

Long run frequency. There exists a ground truth and we have noisy samples of this truth. We attempt to achieve objectivity through statistical tests, confidence intervals, maximum likelihood. Often no probability for single, non-repeating events

## What is the Bayesian approach to probability?

Probability is a degree of belief on a hypothesis given incomplete data. Probability is highly subjective, depends on the current state of knowledge of the agent. Key concepts are: Prior, Posterior and model averaging.

## What is the formula of Bayes’ Theorem?



## What is the technical difference when predicting using frequentist approaches or Bayesian approaches?

* Frequentist: Ground truth physical parameters, prediction based on point estimates, uncertainty quantification via confidence intervals only when desired
* Bayesian:
  + Treat parameters as random variables.
  + Goal is to estimate a distribution given the data.
  + Prediction by averaging

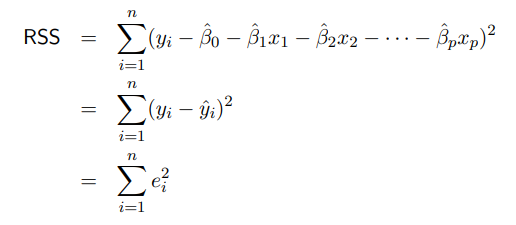
## Correlation is not causation.

Correlation is not causation. We need interventional data.

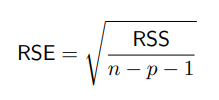
## What is the formula for multiple linear egression?



## What is the formula for the residual sum of squares in multiple regression?



## What is the residual sum of squares defined as for multiple linear regression?

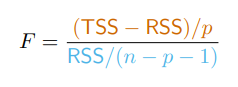


## What is the null and alternative hypothesis for multiple linear regression?

H0: All coefficients are zero.

H1: At least one coefficient is not zero.

## What is the formula for the F-statistic?



## When do the numerator and denominator for the F-statistic become sigma squared?

* If linear model assumptions are correct, the expected value for the denominator becomes sigma2
* If H0 is true, the expected value of the numerator becomes sigma squared
  + F statistic becomes greater the farer away from H0 it is.

## How can we test for significance?

* F-test, p-value depends on p and n

## Shortly describe forward selection, backward and mixed selection.

* Forward selection:
  + Start with 0 features
  + add feature that yields the lowest RSS
  + repeat until termination criterion
* Backward selection
  + Start with full set of features
  + Remove feature with highest p-value
  + Repeat until termination criterion fulfilled
* Mixed selection
  + Start with forward selection, introduce backward steps in between
  + Repeat until convergence

## What are dummy variables? What do they change?

Different encoding for categorical values. Do not change prediction result, but interpretation of coefficients.

## Describe interaction terms.

Interaction terms describe the interaction between two features and treats it as a new feature

## What is the formula for polynomial regression?

Of course extendable

## What are different examples for generalized linear model functions?

Poisson, exponential, logistic (binary), probit

## What is it called when you have a cone shape in your residual plot? Simple fix?

Hereoscedasticity.

Convert response to log, error variance may become constant

## What is an outlier? What is its effect on the regression fit? And on the residual sum of squares?

Data point I where y\_i is far away from y hat \_i

Small effect on regression fit

Notable effect on RSE

## How can you identify outliers? What to do?

Studentized residual plots. Remove outliers or choose another model

## What are high leverage points?

Data points I with extreme x\_i

## What is a criterion for high leverage?



## What is collinearity? What to do?

Highly correlated features, little effect on prediction, possibly large effect on inference

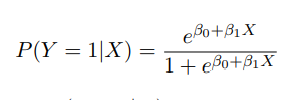
Remove one of the features or average them

# Classification

## What are problems with linear regression when you have categorical features?

How should you encode the features? Which order? How to interpret negative/out of range values?

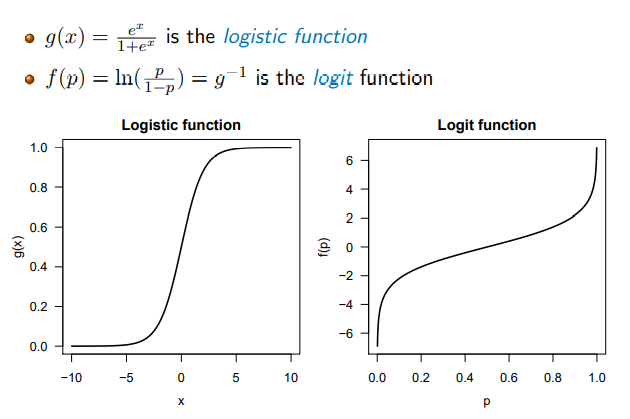
## What is the formula for simple logistic regression?



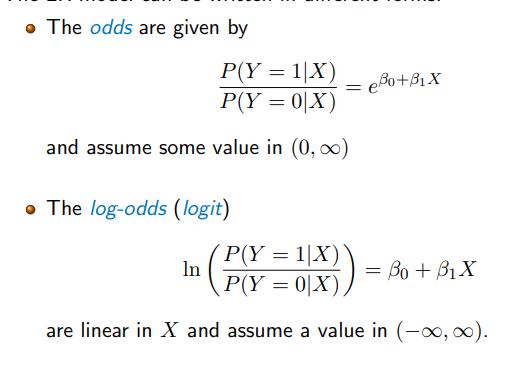
## How does logistic regression work?

Does regression to a probability, assigns probability to sample

## What is the formula difference between the logistic function and the logit function? What do they look like?



## What do they odds and log odds ratio mean in the case of logistic regression? What values do they take?



## What does the mean function look like for logistic regression?

## INSERT HERE A QUESTION ABOUT LIKELIHOOD AND MAXIMUM LIKELIHOOD THAT I DON’T UNDERSTAND YET

## How do you estimate a classification out of probabilities from the logistic regression?

Define a threshold T, assign all data points below threshold 0, all above threshold 1

## INSERT HERE A QUESTION ABOUT MULTINOMIAL LOGISTIC REGRESSION THAT I DON’T UNDERSTAND YET

* 1. K

## Describe discriminative learning approaches.

* Directly solves the task
* Learns to P(Y(,?|?)X) 🡪 joint probability distribution
* Learns a decision boundary
* Model not easily applicable to any other task
* Difficult to optimize exactly 🡪 numerical optimization needed

## Describe generative learning approaches.

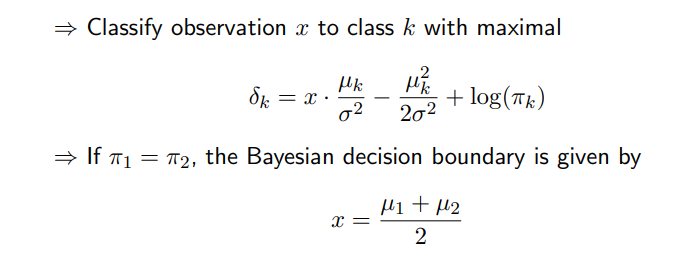
Fit P(X|Y=k) for each class k element Y separately and classify with Bayes’ formula

* Conditional probability distribution
* Each model learns the properties of its class
* Reusable for different tasks (more classes, sampling data)
* Learning P(X|Y) is often easier than learning P(Y|X)

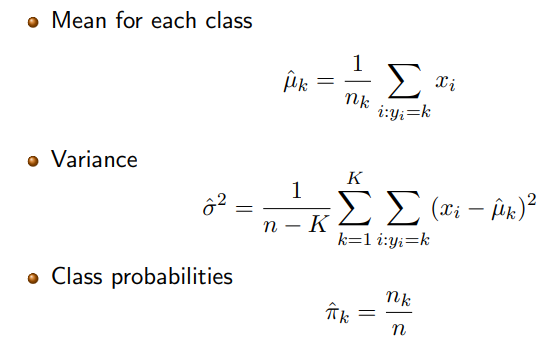
## What is the formula for the gaussian distribution?



## What is the idea behind linear discriminant analysis`?

* Each class function is modelled by a Gaussian distribution
* Each class has its own mean
* All classes share the same variance
* 

## How are the parameters for linear discriminant analysis estimated via the maximum likelihood principle?

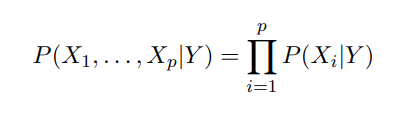


## How does quadratic discriminant analysis differ from linear discriminant analysis?

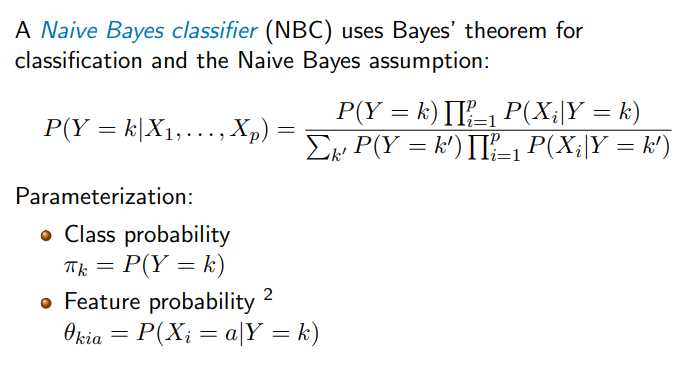
* Class specific covariance matrices sigma
* Quadratic in x
* Quadratic decision boundary

## What are the underlying assumptions for the Naïve Bayes classifier?

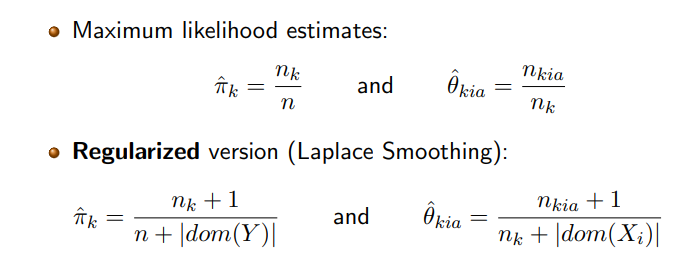
* All features are conditionally independent given class
* Allows different distribution for each feature
* Different variable types



## What is the formula for the naïve bayes classifier?



## What are the estimates for pi and theta for the Naïve bayes classifier?

* 

## What is a key advantage of naïve bayes?

Can cope with missing feature values naturally, just ignore them

## What are some properties of Naïve Bayes?

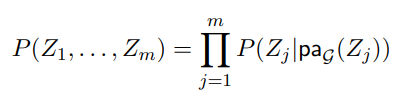
Structure tailored for efficient diagnostics

Unrealistic conditional independence assumptions

Possibly poor calibration

Works in principle also for continuous features

## What is the formula for a Bayesian network based on a directed acyclic graph G?



## Describe a Tree Augmented Naive Bayes.

Y has directed edges to Xx. Nodes X\_i may have at most one other X\_j as an extra parent.

## What are the differences between logistic regression and linear discriminant analysis?

* LDA is better when assumptions are fulfilled
* LDA is better at very small sample sizes
* LDA is faster (closed-form solution)
* Logistic regression makes less assumptions 🡪 more general method
* Dummy variables reasonable for Logistic Regression, not so for Linear discriminant analysis

## When is Naïve Bayes good?

* Missing data
* Entirely or mostly categorical features
* Can handle p >> n problems

## Is discriminative learning always better than generative learning?

* Asymptotically yes – converges to lower error
* In practice not always – slower convergence

## What is the difference between quadratic discriminant analysis and tree augmented bayes and LDA and Naïve Bayes?

* Lower bias, higher variance
* Great if model assumptions are fulfilled
* Otherwise pretty useless

# Performance evaluation

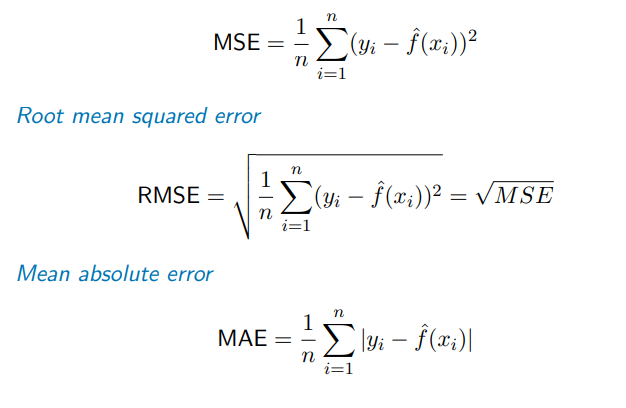
## What are the goals of performance evaluation?

* Model evaluation:
  + Performance in absolute terms
  + or performance against baseline
* Model selection
  + Pick one out a pre-defined set of available models
  + k in kNN, degree in polynomial regression model, logistic regression vs. LDA vs Naïve Bayes
* Stability analysis
  + Stability of parameters estimates or individual predictions

## What are challenges?

* Choosing the correct performance measures
  + MSE for regression? Accuracy for classification?
* No dedicated test set available
  + How to avoid bias?
* Explicit stability measures rely on distributional assumptions
  + May not be satisfied or available at all

## Name three measures used for evaluating regression tasks and their formulas.



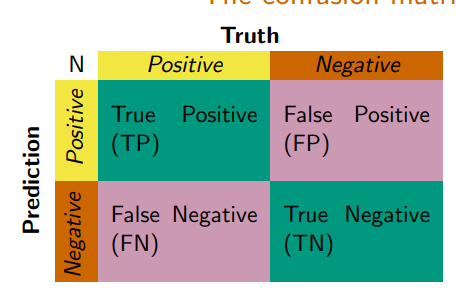
## Which measures penalize larger deviations from ground truth more harshly?

* MSE and RMSE more harshly tan MAE

## What measure is sufficient as a loss function for optimization?

MSE

## What are the components of a confusion matrix?



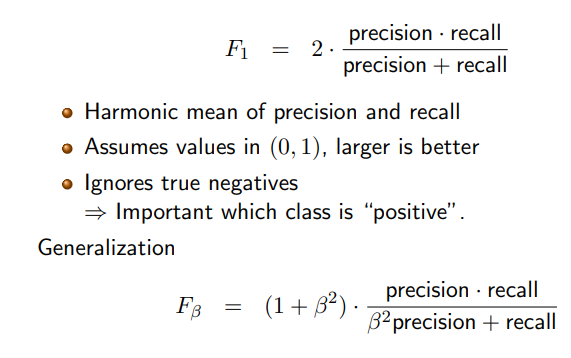
## What are the formulas for sensitivity, specificity and precision?

Precision: TP/(TP+FP)

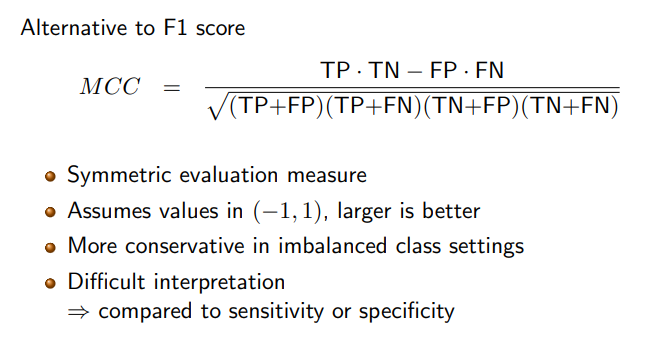
Sensitivity: TP/(TP+FN)

Specificity: TN/(TN+FP)

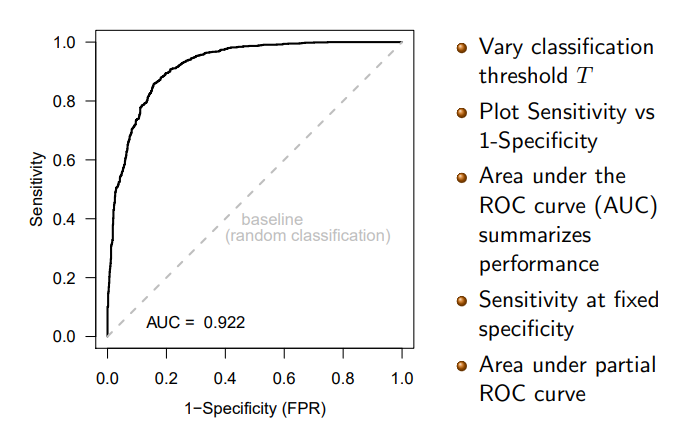
## What is the formula for the F1 score? What are its properties?



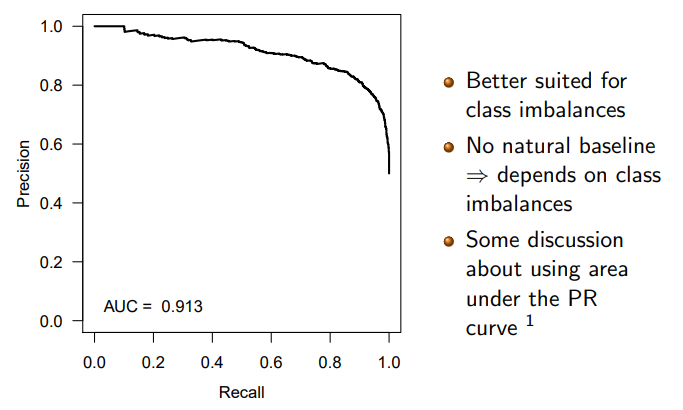
## What is the formula for Matthews’ correlation coefficient? What are its properties?



## How does a receiver operating characteristic (ROC) work? What does it look like? What are the labels on the x and y axis?



## How does a precision-recall curve work? What does it look like? What are the labels on the x and y axis?



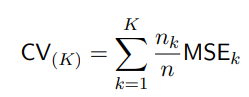
## How can you transform the confusion matrix of a multi classification problem into many binary classification problems?

* Separate into k confusion matrices with the decision “k or not k”. Aggregate results

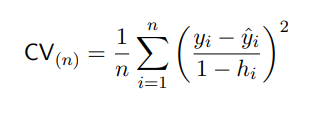
## Describe the holdout approach for data set splitting into training and test set. What are its problems?

* Split data set randomly
* Ratio 1:1 or 2:1
* Problems:
  + Error dependent on sample size
    - Reducing the training data overestimates the error
    - Affects model assessment but also model selection
    - There might be high variance in error estimates

## Describe the process off K-fold cross validation.

* Split the data into K parts
* For testing on part k, train on all data except part k
* Use each part k as test set once
* Average results
* 

## Describe leave-one-out cross validation. What is the analytical solution for linear/polynomial regression?

* Always leave on sample out and test model on that
* No randomness involved
* Many model fits needed
* 

## What should you consider when choosing k?

* Small k yields an unreliable average
* Large k is computationally more demanding
* Small k introduces bias due to reduced sample size
* Large k makes all learned models highly correlated 🡪 higher variance

## What is pooling in regards to performance measures?

* Aggregate performance measures obtained on each fold, e.g. by weighted averaging
* Yields standard error over the different folds

## What is pooling in regards to the original predictions?

* Store raw prediction from each fold
* Concatenate them into one vector
* Compare to ground truth and compute one performance measure (without standard error)

## Describe repeated holdout. What is the difference to k-fold CV?

* Split data into training and test at a given ratio
* Repeat many times
* Average results
* Advantages:
  + Number of iterations is independent of the training data size
  + More iterations can be added if needed
* Disadvantages
  + Test data points are not chosen equally often
  + More splits to store for reproducibility

## What are other repetition strategies and how do they work?

* Repeated k-fold cross validation
  + Multiple splits
  + Average results
* Stratified cross validation/holdout
  + Ensures that response has same distribution in all splits
  + Good idea with high class imbalance

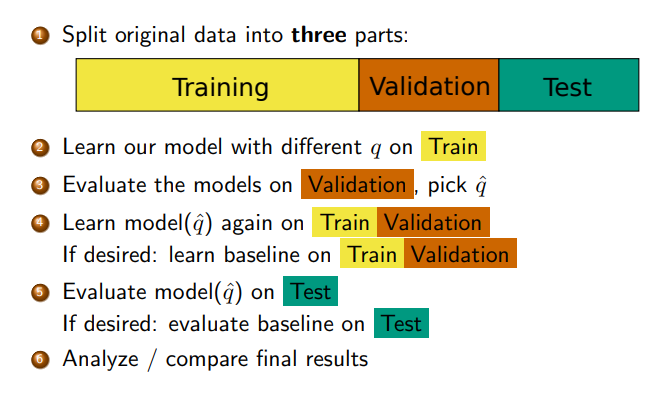
## You have obtained the average errors and standard errors for your evaluation strategy. Which model should you choose?

1. Lowest standard error
2. One standard error rule: choose most parsimonious model that is within one standard error of the optimal model

## What is the problem with performing simple hold-out evaluation for tuning a hyperparameter hat q?

- if a model yields a better error than our baseline error, we cannot claim that it is better, because we implicitly use the test data set for estimating hat q.

## Describe (with illustration) the process of nested cross validation.



## What would be the ideal process or quantifying uncertainty of an estimate? Why is that not possible?

Sample multiple data sets from the true distribution of the data

Estimate the parameter of interest from each data set

Compute standard errors from all computed estimates

* Problem: The true distribution of the data is not known

## What is the idea and process of bootstrapping?

-Treat the given distribution as a surrogate for the true data distribution

Randomly sample a data set by drawing n data points from Z wit replacement

Estimate hat a from samples

Compute uncertainty measure from all hat a’s

# Regularization

## What are the advantages of regularizing linear models?

* Ordinary least squares cannot be used with no. features > no. samples
* Low variance of linear regression becomes high when no. samples not much bigger than no. features 🡪 Danger of overfitting
* Higher interpretability

## Name three methods for regularization and briefly describe them.

* Subset selection
  + Identify a subset of p features that are most predictive
  + Fit LR model with least squares on this subset
* Shrinkage methods
  + Fit a model on all p features
  + Modify objective function to penalize large coefficients 🡪 coefficients are shrunk towards zero
* Dimension reduction
  + Project p predictors into an M-dimensional subspace
  + Fit LR model with least squares in this subspace

## Name three methods one can use to choose between models of different complexity (number of features).

* + 1. - likelihood ratio test (requires nested models)
    2. - cross-validation (expensive)
    3. - model selection criterion (cheap): corrects training fit for model complexity

## Describe the process of forward stepwise selection.

* Model M0  contains 0 features
* For k = 1,2,…,p:
  + Test all possible models that increase Mk-1 by one feature
  + Pick the best model according to R2 🡪 Mk
* select the best overall model among M­0,…,Mp according to model selection criterion

## What are advantages and disadvantages of forward stepwise selection?

* Cheap
* Greedy
* If number of features is bigger than number of samples, it selects from models M0, …, Mn

## What is the complexity of forward stepwise selection?

* Quadratic complexity: 1 + p\*(p+1)/2

## What is the process of backward stepwise selection?

* Model Mp contains p features
* For k = p, p-1, …, 1:
  + Test all possible models that decrease Mk by one features
  + Pick the best model according to R2 🡪 Mk-1
* Select the best overall model among M0,…,Mp according to some model selection criterion

## What are the complexity and sample requirements of backward stepwise selection?

* Complexity like forward stepwise selection
* No. features <= no simples (to fit Mp)
* Can be combined with forward selection to a hybrid approach

## What is the basic formula for shrinkage methods?

* RSS + g(ß)
* Where g maps from Rp to R>=0 and is function of coefficients (excluding intercept) and is called shrinkage penalty

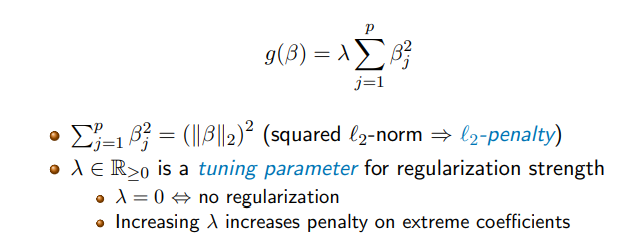
## Why do shrinkage methods work?

* Coefficients will be smaller on average than in Ordinary Least squares estimate
* An optimal model has residual sum of squares >= OLS estimate

## Name some variants for shrinkage methods

* Ridge regression, lasso, elastic net

## What formula is used for g in ridge regression? What do the parameters mean?

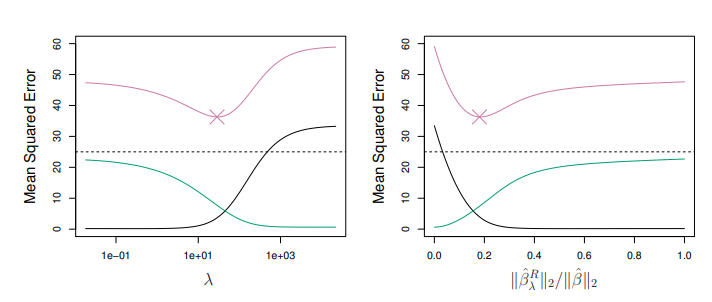


## What are the main disadvantages of ridge regression?



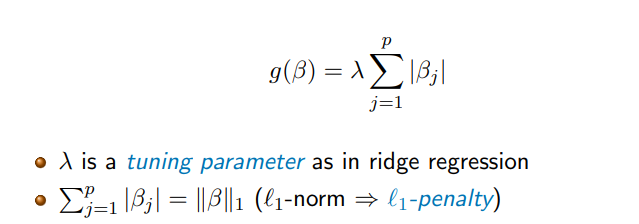
* Shrinks many coefficients towards zero, but not exactly zero, which makes interpretation difficult.
* Differences in measurement method are significant (e.g. m and cm), need to be standardized

## Draw an example for the course of variance, squared bias and test MSE on real data for lambda and l2-penalty.



Green: variance, black: squared bias, purple: test mse

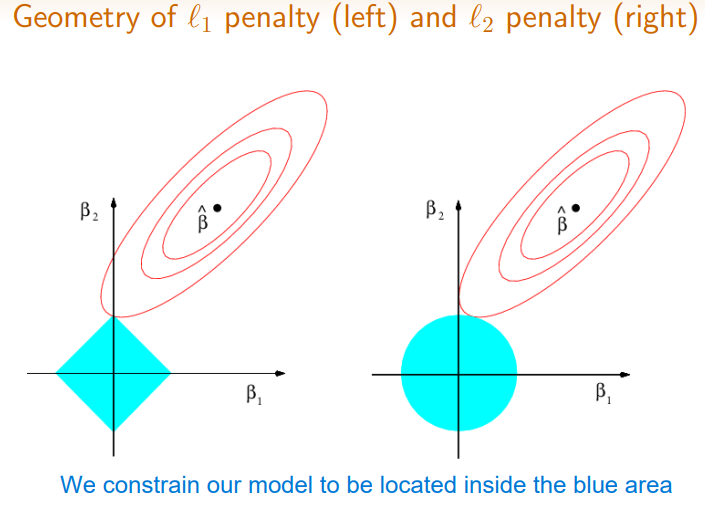
## Describe the formula for lasso. What do the parameters mean?



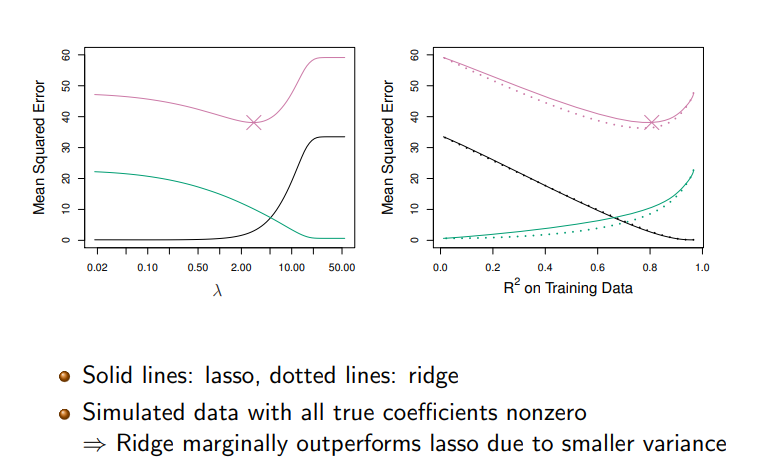
## Conceptually, what is the main difference between Lasso and Ridge Regression and Subset selection?

* Many features are ignored entirely in lasso
* Lasso is good for interpretability
* Lasso is more efficient than subset selection

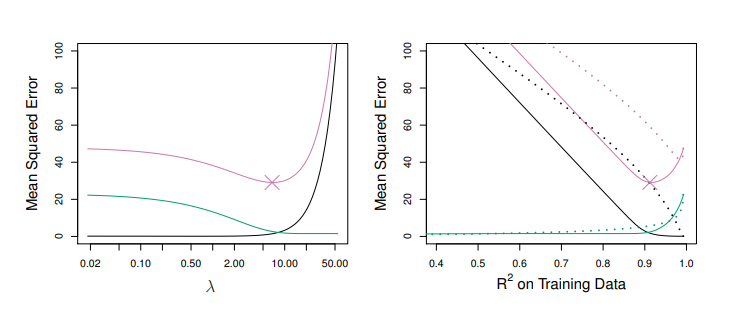
## What are the geometries of the l1 and the l2 penalties?



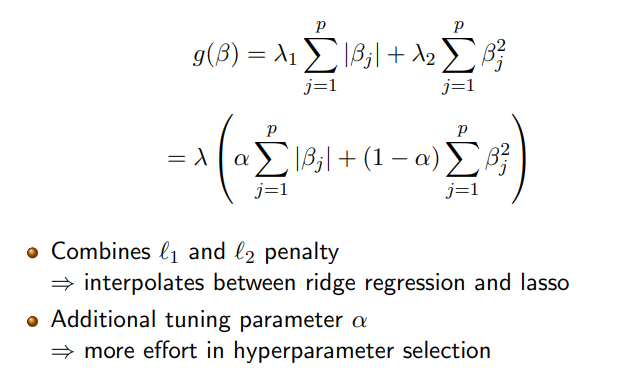
## What does it look like when you plot lambda and R2 on training data by mean squared error for Ridge regression and lasso?



## What does it look like when you plot lambda and R2 on training data by mean squared error for Ridge regression and lasso and some coefficients are exactly zero?



## What is the formula for the penalty term for Elastic net?

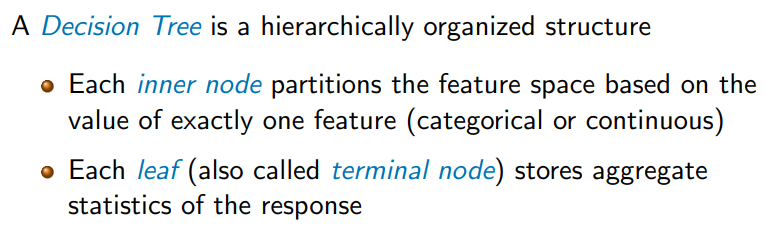


## Describe the process of parameter tuning.

* Choose a grid of lambda values
* Compute cross validation error for all possible values
* Determine lambda hat by choosing the model with minimal cross validation error
* Refit the model with lambda hat on entire training data

# Decision Trees

## Describe the structure and parts of decision trees.



## What are the two kinds of decision trees? What do the leaves store?

* Classification tree with categorical response
  + Leaves store e.g. counts for every category
* Regression tree with continuous response
  + Leaves store e.g. mean and variance

## What are the advantages and disadvantages of decision trees?

Advantages

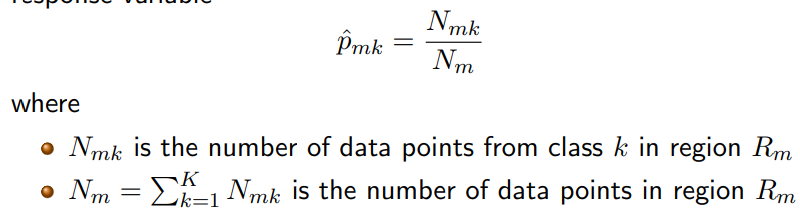
* Easy to explain
* Mirror human decision-making
* Can easily handle categorical predictors

Disadvantages

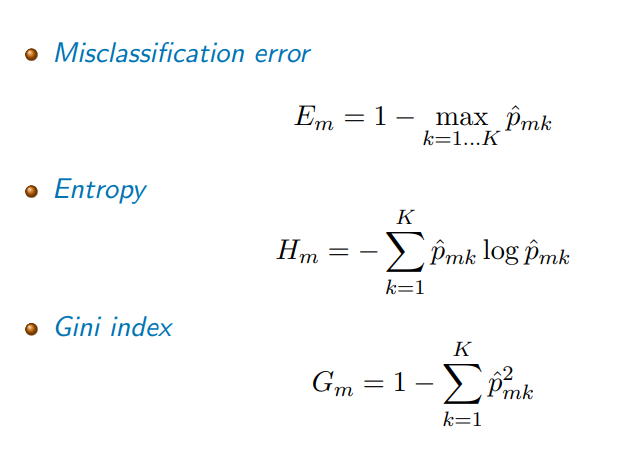
* Limited predictive performance
* Non-robust, tend to overfit

## How can one evaluate the quality of a region?

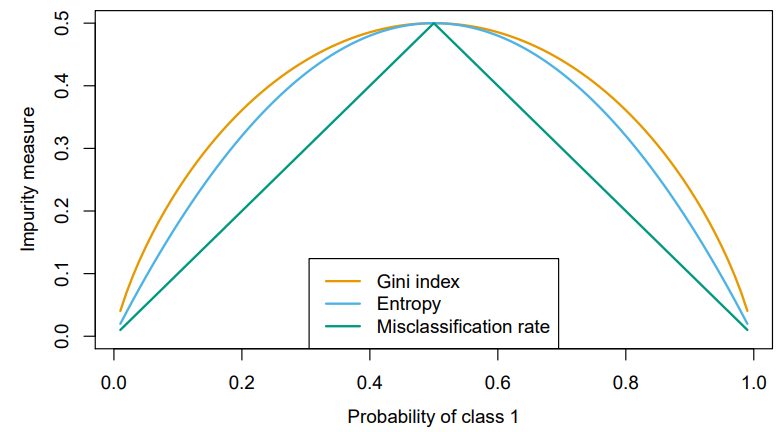
* Quality depends only on the distribution of the response variable



## Name some impurity measures for decision trees. What are their formulas?



## What do the different impurity measures have a course like depending on the probability of class?



## What approaches are there to control tree complexity?

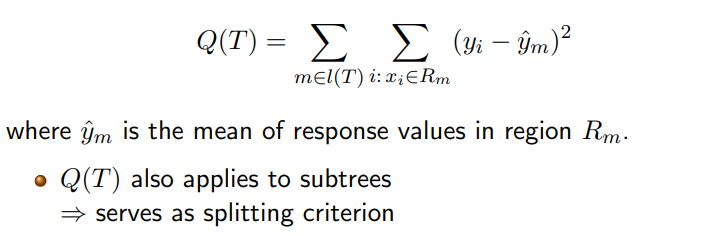
* Limit number of splits
* Limit maximal depth
* Enforce minimal leaf size
* Require impurity difference to be over a certain threshold for making a split
* Prune a fully grown tree from the bottom up

## When fitting a regression tree instead of a classification tree, how is it done?

* Different splitting criteria
* Different pruning
* Different prediction

## How is the quality of tree T defined?

By the residual sum of squares.



## What are characteristics of this quality measure?

* Characteristics of different leaves are proportional
* No meaningful concept of purity 🡪 recursion terminates when a region contains sufficiently few data points

# Ensemble Learning

## What are the two kinds of ensemble learnings?

* Homogeneous ensemble: same base learning algorithm
* Heterogeneous ensemble: Different base learning algorithms (rare)

## What are the two types of ensemble methods?

* Parallel ensembles: independent base learners
  + Bagging, random forest
* Sequential ensemble: base learners depend on each other
  + Boosting

## How does the average variance of f models trained on B data sets change in respect to the original variance?

## 

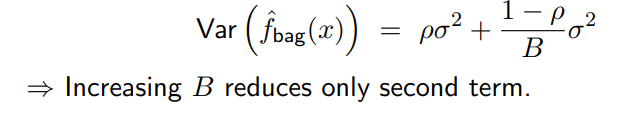
## How does bagging work? What kind of ensemble does it produce?

* We only have on training data set Z
* Using bootstrapping, we construct Z\*1 ,…,Z\*B
* We fit one model on each data set, each with variance sigma^2
* Then we predict
* It produces a parallel ensemble

## What are the limits of bagging?

* Limited variance reduction
* Bootstrap estimates are only identically distributed, not independent (subsamples of same data set)
* The single data sets have pairwise correlation rho

## What is the variance of bagging? What does increasing B do?



## Explain how to out of bag error.

* For each model, use subset of training data (rest is out-of-bag observations)
* For predicting hat y\_i , use only models where sample is out of bag
* Average on predictions or do a majority vote
* Repeat for all n data points

## What is the idea of random forests?

Before each split, randomly select m features as candidates for splitting. Find the best splitting option among these.

## What is the aim of random forests?

Decorrelate the trees (reduce rho)

## Describe the process to learn a random forest.

For b = 1,…,B:

Draw bootstrap sample Z\* of size N from training data

Grow decision tree Tb on Z\* by recursively applying the following steps

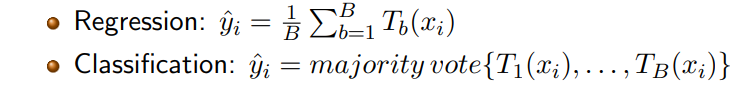
Select m out of p features at random

Pick best variable/split among the m features

Split the node

Output ensemble of trees

## What are the two formulas that a sample hat y\_i is predicted by by regression and classification?



## What are the two most important hyperparameters and their default values?

* Number of trees grown (500)
* Mtry (lower bound square root p), mtry (lower bound p/3)

## When should you stop splitting for regression trees?

* When leaf has 5 or less data points in regression

## Is boosting building a sequential or a parallel ensemble?

Sequential

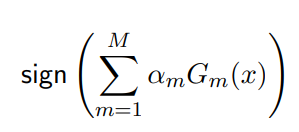
## Name some characteristics of boosting.

* Only a single data set, no bootstrapping
* Base learners are simpler than for bagging (even just one decision)
* More models reduce bias

## Describe the process of adaptive boosting.

* Each training data point gets a weight
* For each model:
  + Learn model G\_m on weighted data
  + Compute training error err\_m
  + Update weights based on err\_m (increase weights of misclassified data points)
* Combine the M models by a weighted majority vote

## What is the formula for the weighted majority vote in Boosting and what are its parameters?



Alpha\_m depends on err\_m

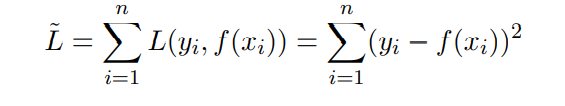
M models

G\_m(x) is model prediction of model m

## How does AdaBoost for regression work?

Model is fitted on data, the residuals are computed and then the next model is computed in respect to those residuals. Then the two models are combined.

## What function do we minimize in gradient bosting?



## What differences are there between gradient boosting and XGBoost?

* More regularization options: l1, l2, number of leaves
* More randomization options

## How can one interpret variable importance in a random forest?

* To many trees, so we have:
  + Mean decrease in impurity (decrease in Gini index when splitting according to a feature of interest)
  + Sum it over all splits with that feature in the tree (weighted by probability of reaching the node)
  + Average over all trees
* Mean decrease in accuracy
  + Permute values of feature of interest in OOB data
  + Compute decrease in accuracy

## What is one word of caution when interpreting variable importance?

* Forest are biased to splitting variables with lots of splitting points (continuous, categorical with many different categories)

## How could one visualize the relative importance between two features?

* Partial dependence plots

# Support Vector Machines

## How do support vector machines separate two classes? How do they achieve nonlinearity?

Separate two classes with an optimal hyperplane. Nonlinearity by feature transformation via kernels.

## What two special cases of SVMs did we have in lecture?

Maximal margin classifier 🡪 Linear SVM with hard margin

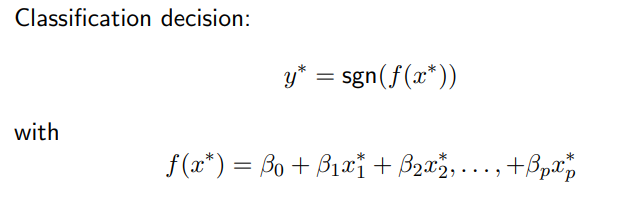
Support vector classifier 🡪 Linear SVM with soft margin

## What is a hyperplane?

A hyperplane is a flat affine subspace of dimension p-1 in a p-dimensional feature space.

## How does one come to a decision regarding y\_i with a hyperplane?

The hyperplane is positive when y\_i is 1, the hyperplane is negative if y\_i is -1.



## How do you find an optimal separating hyperplane?

Compute distance from each data point to the separating hyperplane

Choose hyperplane that maximizes the margin

Data points with minimal margin are support vectors

## How can you solve for the maximal separating hyperplane?

Lagrange duality

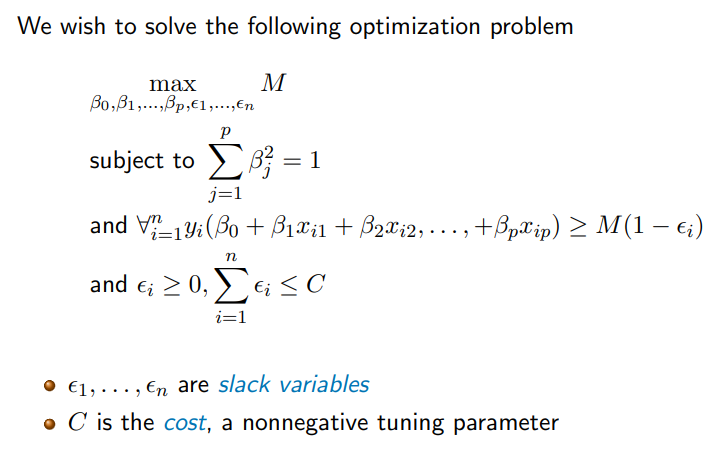
## What are problems of the maximal margin classifier?

* Can’t find separating hyperplane for non-separable data
* Very sensitive to individual data points

## What is a soft margin in a support vector machine?

* A soft margin does not attempt to perfectly separate ethe classes
* It allows for a few misclassifications
* Data points violating the margin are also support vectors
* Resulting classifier is called soft margin classifier/support vector classifier

## Describe the optimization problem for the support vector classifier.



## What are slack variables and what are the different cases for slack variables?

Slack variables eta\_i encode location of data point i w.r.t. to hyperplane and margin

* Eta\_i > 1 🡪 i on wrong side 🡪 support vector
* Eta\_i > 0 🡪 i violates the margin –> support vector
* Eta\_i = 0 🡪 i on correct side 🡪 no support vector

## What does it mean if C has a high/low value? How is C chosen?

High cost means low variance, high potential bias and many support vectors.

Low cost means high variance, low bias and few support vectors.

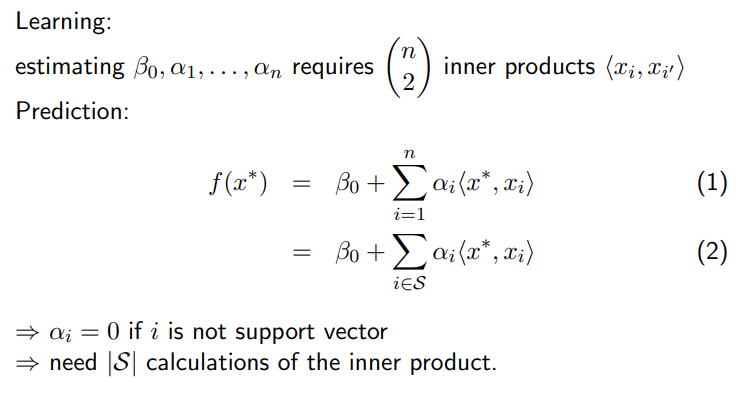
* Choose C via crossvalidation.

## What is one solution to fit SVCs on non-linear data without using kernels?

* Enlarge the feature space
  + Yields non-lienar decision boundary
  + Higher degress & interaction terms also possible
* Problem number of features grows, computation becomes more expensive

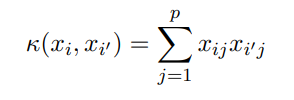
## How does a general Support Vector Machine work?

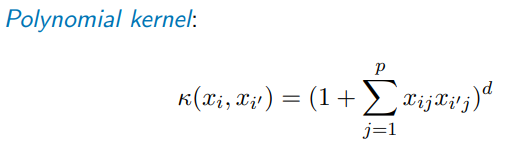
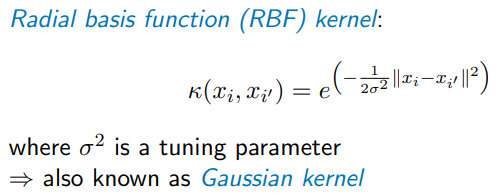
Replaces general hyperplane formula with inner product. Support vector machine replaces inner product by a kernel function kappa(x\_i, x\_i’)



## Name some kernels and reproduce their formulas.

* Linear kernel:



* 
* 

## What is the advantage of using kernels vs. an enlarged feature space?

* Computationally more efficient, only needs to compute the kernel matrix
* Some kernels may even represent infinite-dimensional feature space

## What should you do if you have more than two classes?

* One-vs-one classification:
  + Compare kth class to k’th class
  + build (k over 2) SVMs for all pairwise combinations
  + classify x\* according to each SVM
  + predict the class that was chosen most often
* One-vs-all classification
  + Compare kth class to all other classes
    - Does data point I belong to class k or not
    - Build K SVMs
  + Classify x\* to class k which maximizes



# Clustering

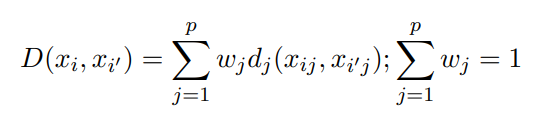
## Is Clustering a supervised or unsupervised learning approach?

Unsupervised

## What are the goals of cluster analysis?

* Group objects into clusters
* Obtain an additional hierarchy of clusters
* Find out whether different subgroups exist in the data at all

## What is dissimilarity (with formula)?



P features, weights w, distance function d.

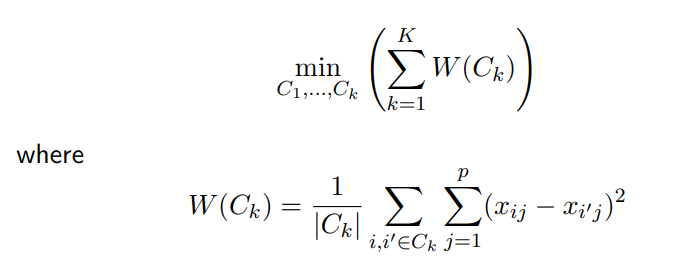
## What are common dissimilarity functions?

* Continuous: Squared Euclidian distance
* Categorical: We need dissimilarity matrix

## What are the ideas of k-means?

* Partition the data into K non-overlapping clusters
* Choose clusters so that within-cluster variation is minimized

## What are the minimization problem formulas for k-means?



## Why is finding the globally optimal solution for k-means hard?

* There are nk different partitionings of n items into K clusters

## What is the algorithm for finding a locally optimal solution?

* Randomly assign each data point to one cluster
* Iterate until cluster assignments do not change anymore:
  + For each C\_k compute the centroid
  + Assign each data point to its closest cluster center 🡪 Squared Euclidean distance to cluster centroid

## What is the k-means partitioning called?

Voronoi tessalation

## What is a common problem for k-means classification?

Label switching

Locally optimal solutions

## How can one choose the optimal K?

Log-plot of W\_k. Choose the k with the kink in the log plot.

## What is the process for a Gaussian mixture model?

* Choose K multivariate Gaussian distributions
* A hidden variables u\_i indicates cluster of x\_i
  + Seek cluster membership probability w\_ik
* Learn via EM algorithm by iteratively updating
  + Model parameters theta
  + All w\_ik
* When variance is 0, algorithm coincides with K-means
* Othherwise it returns soft cluster assignments

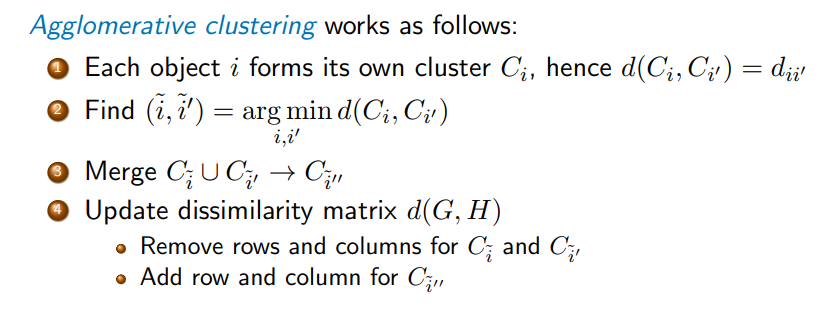
## Explain K-medoids.

* K-medoids allows arbitrary distance measures
  + Chooses medoid I element of C\_k as cluster representative
    - Data point that is closest to all other data points in C\_k
  + Needs O(n2k ) computations (instead of O(n\_k) for the centroid)
    - Faster versions for the average case

## What is the idea of hierarchical clustering?

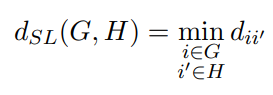
* No need to specify the numbers of desired clusters
* Build a hierarchy, represented by a dendrogram
* Cutting the dendrogram at a specific height yields clusters 🡪 one tree may yield any number of clusters

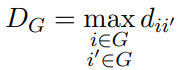
## What is the process of agglomerative clustering?



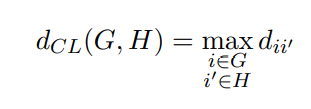
If more than one cluster left, return to step 2

## What is the formula and what are the characteristics of single linkage?



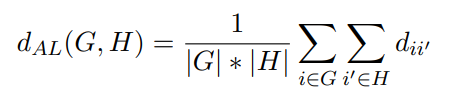
* Single linkage yields chaining: clusters are merged if they marginally touch each other
* Clusters have large diameter:  🡪 violates compactness property

## What is the formula and what are the characteristics of complete linkage?



* Complete linkage produces compact clusters, small diameter
* Element of cluster D can be closer to element i’ of cluster H than to element i’ of cluster D 🡪 violates closeness property

## What is the formula and what are the characteristics of average linkage?

* 
* Compromise between single and complete linkage
* Result depends on scale of dii’ 🡪 even monotone transformations might change resulting dendrogram

## What is a particular use case for agglomerative clustering?

* Phylogenetic inference

## What is the process for reconstructing a phylogenetic tree?

Objects: DNA sequences from different organisms

Procedure:

* Compute pairwise alignments
* calculate distances using a model of molecular evolution
* run hierarchical clustering (e.g. UPGMA)

## What is the process of divisive clustering?

* Start with all data points contained in one cluster
* Split cluster into two subclusters based on cluster dissimilarity
* Repeat the process recursively
* Terminate when all cluster have size

## Why is divisive clustering rarely used in practice?

* Splitting cluster into subclusters based on cluster dissimilarity may take O(2n) time

# Dimension reduction

## What is the goal of dimensionality reduction?

Project data onto a lower-dimensional space while retaining important properties of the data

## What is the motivation for dimension reduction?

Data preprocessing for supervised learning

* Computational efficiency
* Noise reduction
* Improving bias-variance tradeoff

Data preprocessing for clustering

* Computational efficiency
* Noise reduction

Data visualisation

* Projection from 3D to 2D
* “eyeball clustering”

## Name dimension reduction methods

* Principal component analysis
* t-stochastic neighbor embedding
* non-negative matrix factorization
* linear discriminant analysis
* uniform manifold approximation and projection
* autoencoder

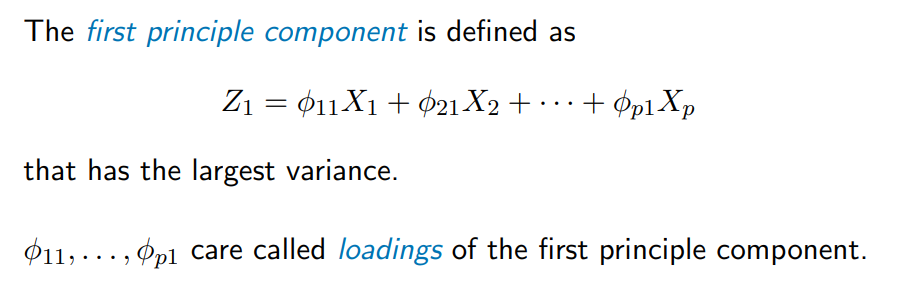
## What is the general idea of principal component analysis?

* Linear mapping to lower-dimensional space
* Variance in lower-dimensional space is maximized
* Basis vectors of that space are called principal components

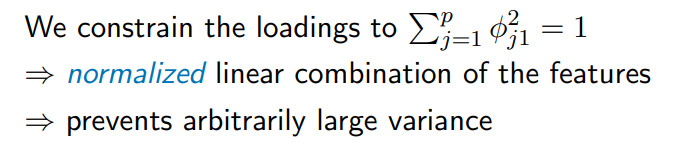
## What preprocessing should be done to the data before PCA?

* Mean-centering
* Often recommended: scaling to standard deviation of one

## What is the formula of the first principal components? What are the factors called?

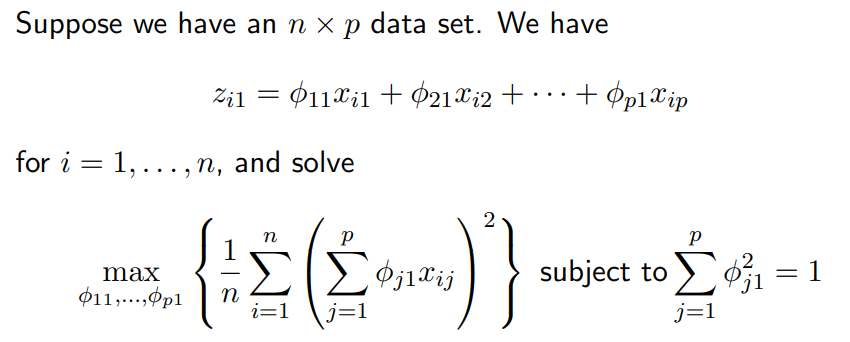


## What value constraint do the loadings have?



## How do we solve for the first principal component? What formula do we maximize?

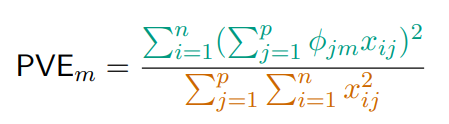
We solve for it via Eigen decomposition



## What conditions does the second principle component need to fulfil?

* Linear combination of Xi, … , Xp
* Uncorrelated with Z\_1 (orthogonal)
* Maximizes the variance

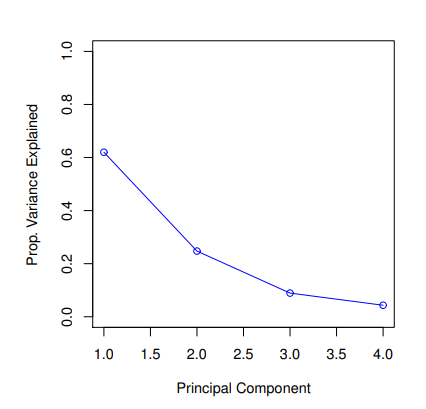
## What is the formula for the proportion of variance explained?



Green: variance explained by component m. Orange: Total variance.

## What does a Scree plot show? Draw an example.

Proportion of variance explained vs. M (no of principal components)



## What is the main difference between PCA and t-SNE?

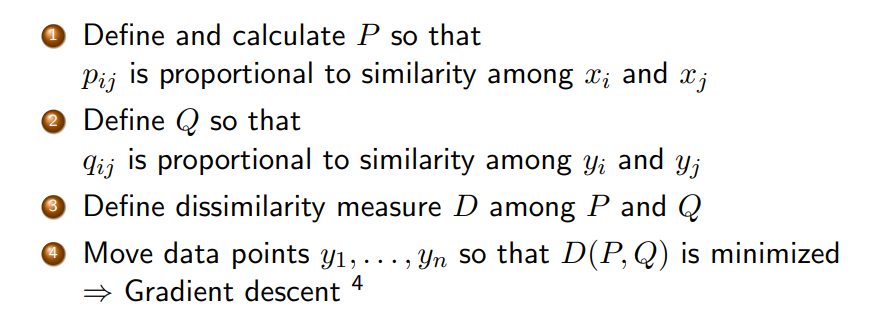
PCA preserves large pairwise distances through variance maximization 🡪 sacrifices small pairwise distances

T-SNE preservs local structure (small pairwise distances)

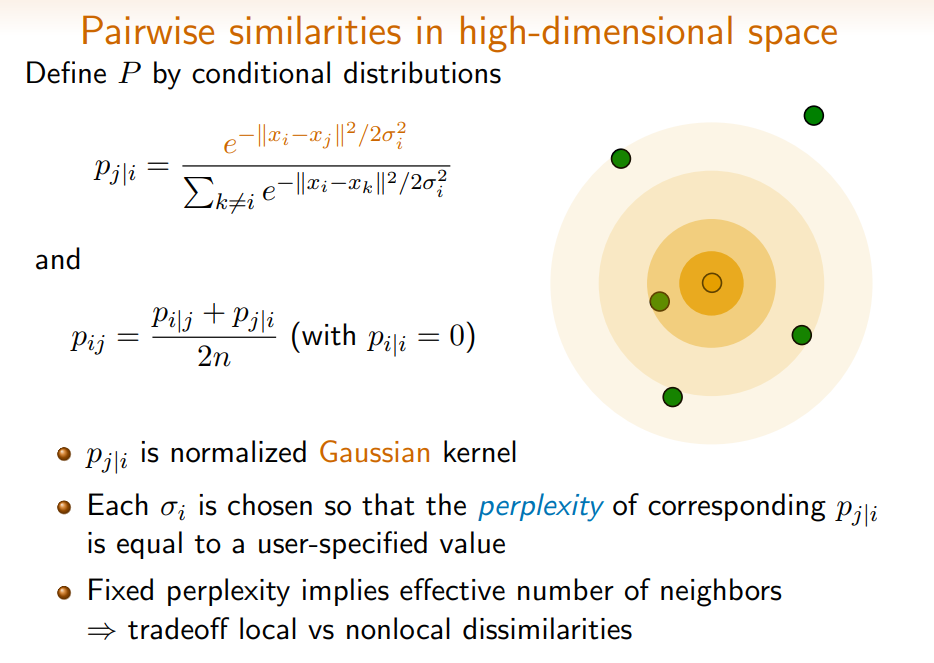
## What are other approaches that preserve small pairwise distances?

* Isomap
* Local lienar embedding
* Sammon mapping

## What are the parameters of t-SNE and what is the process for its optimization?

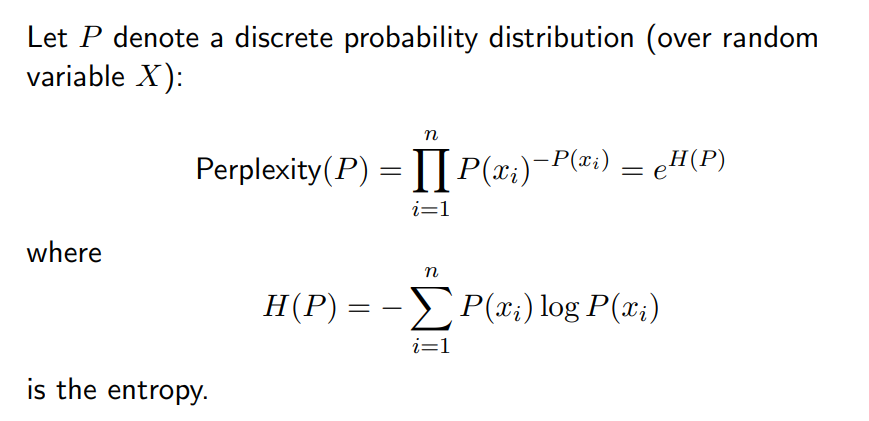


## How should you choose P?

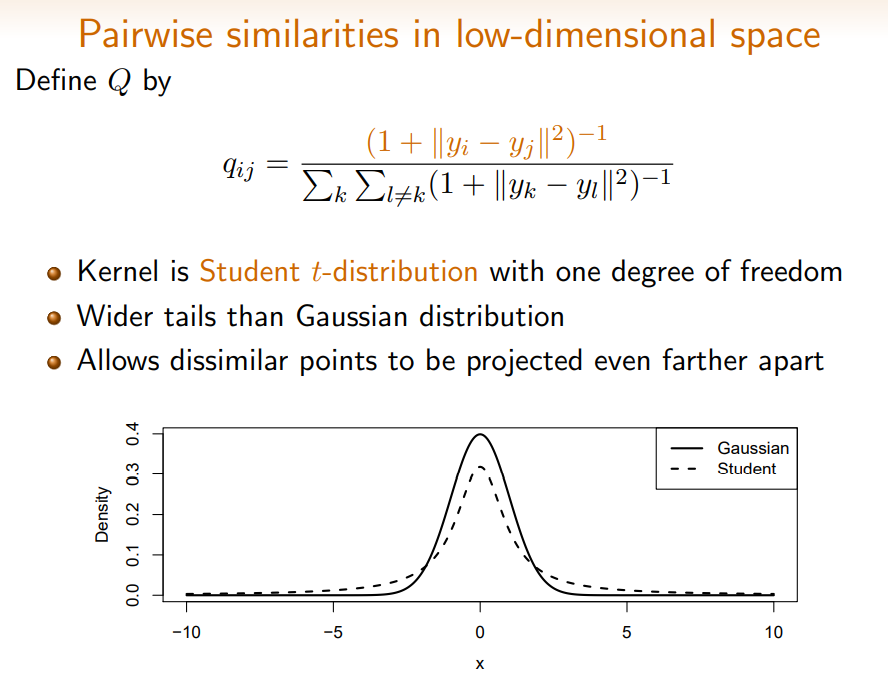


## What is perplexity?

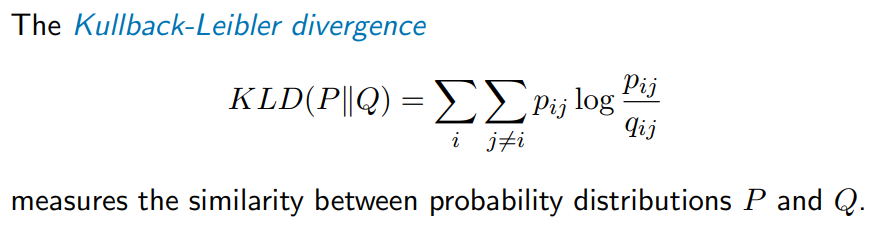
Basically entropy



## What is Q?



## What is the formula of the Kullback-Leibler divergence? What does it measure?



## How is the Kullback-Leibler divergence asymmetric? What does that mean for different values for p and q?

Large p\_ij, small q\_ij 🡪 Large penalty

Small p\_ij, large q\_ij 🡪 small penalty