TTK4260
Multivariate data analysis and machine learning
by Jan-Øivind Lima

Notes from exam preparation

Contents

§1. Introductions and motivations (Lecture 1)

The overarching workflow for data analysis is seen below:

Example 1.1 (The main piece of the data analysis workflow).
- Consider the best approach given the data and objective
- Split the data into training/validation/test
- Decide on preprocessing of the training data; save this procedure for validation and test sets
- Proposing some alternative models
- Estimating the model parameters
- Choosing the best model structure
- Model interpretation given background knowledge (theory, literature, own experience)
- Identify possible outliers and decide whether to take them out or not
- Validating the chosen model(s)
- Estimating the performance of the model, including parameter uncertainties

For modeling there are several strategies to employ, here are a few of them.

- Mechanistic models / physicas based models: based on theory, first principles and/or domain knowledge (deduction)
- Data-driven models: Models based on carefully designed experiemntal (numerical of physics) data. They can be black box, grey box or white box (induction)
- Meta Modelling / Hybrid Modelling: A combination of the above two.

Some desired characteristics in the modeling approach are:
- Generalizability / Robustness
- Trustworthiness / Transparent / Explainable
- Computational efficiency yet accurate
- Dynamically adapting and evolving

§1.1. Dimensionality of a model

Most systems are observed iwth more sensors/variables than the actual underlying dimensinality(rank). Methods based on latent varibales will summarize the information in terms of "super varibles" and reveal the individual variable's contribution to these. The rank of a system can mean different things:
- The **numerical** rank (This will for all real data be the smallest dimension of the data table)

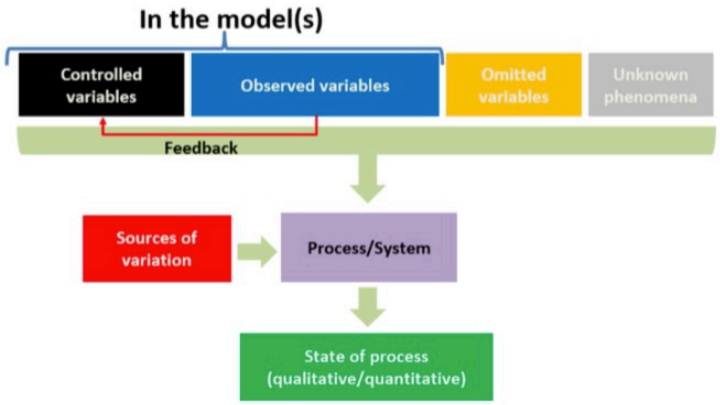


Figure 1: Variables in a system

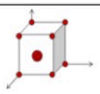
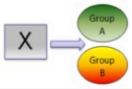
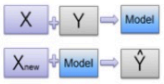
Tool	Symbol	Purpose
Design of Experiments (DoE) Full and Fractional Factorial Designs, Optimisation Designs		Gain maximum information from a minimum of experimental effort
Exploratory Data Analysis (EDA)		Find important sample groups or variable relationships in large data sets. Define classes.
Classification/Discrimination		Classify samples. Identify raw materials. Is the process under control (MSPC)?
Regression and Prediction		Predict quality for new samples and/or find out which variables that most influence a process

Figure 2: Tools used in this course

- The **statistical** rank (found by some statistical criterion, e.g. the chosen validation scheme or maximum likelyhood)
- The **application** specific rank based on experience and interpretation

§1.2. Redundancy and selectivity

Most sensors are not measuring directly the inherent state or property of a system. Although individual sensors are not selectie, ,we can still use them for observin the system:

- **Qualitative**: Is the system under control?
- **Quantitative**: What is the water content in this rock on Mars?

§1.3. Typical pitfalls for valitation of classification or prediction models

- No independent test set
- Replicates in training and test sets respectively
- Use information in the model that is not present during prediction (for example using time as a variable)
- Selection bias (especially when tuning hyper-parameters)
- Extending sample sets with simulated samples to e.g. handle skewed distributions, then split in training/test
- Feature selection on all data, then split in training/test (see above)
- **And what is not often mentioned: Random training/test when the samples must be stratified due to meta-information**

§2. Principal Component Analysis (PCA) introduction (Lecture 2)

PCA is the "mother" of all exploratory multivariate methods, below are some examples demonstrating the need/utility of the exploratory data analysis.

Example 2.1.

- Pattern recognition
- Dimensionality reduction
- Clustering and classification
- Outlier detection
- Condition monitoring and predictive maintenance, e.g. of wind turbines
- Denoising
- Data imputation
- Speed up training of machine learning models (using latent variables)

§3. Least squares estimators (Lecture 3)

Indtroductory definitions:

- $y \in \mathcal{Y}$: random variable (r.v.) from a probability distribution P:

$$y \sim P(\theta) \qquad \theta \in \Theta$$

(Example: $y \sim \mathcal{N}(m, \sigma^2)$)

- $y_1, \ldots, y_N = \boldsymbol{y} \in \mathcal{Y}^N$: sampled data

Definition (statistic)

$$\phi(\boldsymbol{y}) : \mathcal{Y}^N \mapsto \mathbb{R}^M$$

is a statistic of $\boldsymbol{y}$ if

- *it is measurable*
- *it does not depend on θ*

Figure 3: Introductory definitions of least squares estimators

§3.1. Estimator

Estimator of θ

- $y \in \mathcal{Y}$: random variable (r.v.) from a probability distribution P:

$$y \sim P(\theta) \qquad \theta \in \Theta$$

- $y_1, \ldots, y_N = \boldsymbol{y} \in \mathcal{Y}^N$: sampled data

Definition (estimator (of θ))

any statistic of $\boldsymbol{y}$ with co-domain equal to Θ, i.e., any

$$\phi(\boldsymbol{y}) : \mathcal{Y}^N \mapsto \Theta$$

that

- *is measurable*
- *does not depend on θ*

Figure 4: Estimator

§3.2. Least sqares form geometrical interpretations

First the assumptions:

data generation model: $y_t = f(u_t; \theta) + v_t$

dataset: $\mathcal{D} = \{(u_t, y_t)\}_{t=1,\ldots,N}$

hypothesis space: $\theta \in \Theta$

Problem: *find a* $\widehat{\theta} \in \Theta$ *that "best explains"* $\mathcal{D}$

Figure 5: Assupptions for model.

This leads to the geometrical interpretation:

$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$ fixed $\quad \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{bmatrix}$ fixed $\quad \begin{bmatrix} f(u_1;\theta) \\ f(u_2;\theta) \\ f(u_3;\theta) \\ \vdots \end{bmatrix}$ manifold in θ

example with $\theta \in \mathbb{R}^2$:

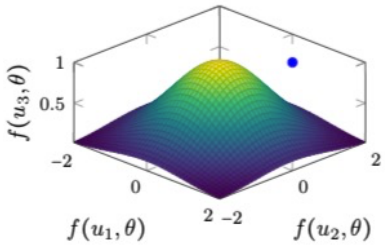


Figure 6: Geometrical interpretation of least squares

§4. Statistical performance indexes (Lecture 4)

There are three measures of central tendencies, they are **mode**, **median** and **mean**.

Example 4.1.
- **Mode**: most commonly observed value
- **Median**: midpoint
- **Mean**: average

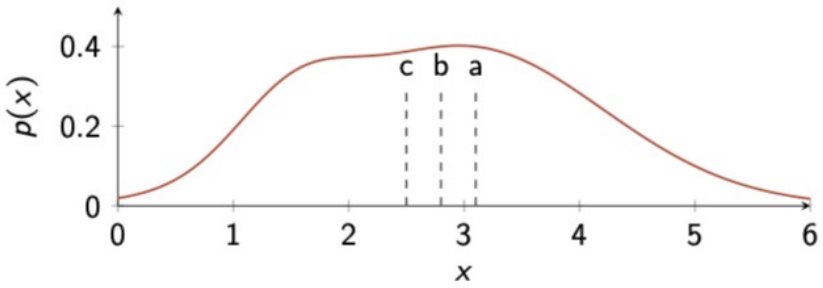


Figure 7: Central tendency

Measures of dispersion are **range**, **variance** and **standard deviation**.

Example 4.2.
- **Range**: maximum - minimum
- **Variance**: mean of the squared differences between the elements of a dataset and their mean
- **Standard deviation**: square root of the variance

Measures of association are **covariance** and **correlation**.

Example 4.3.
- **Covariance**: measure of the joint variability of two random variables
- **Correlation**: measure of the strength and direction of the linear relationship between two random variables

Some other measures are **skewness** and **kurtosis**:

Example 4.4.
- **Skewness**: how symmetric a probability distribution is
- **Kurtosis**: how tailed a probability distribution is

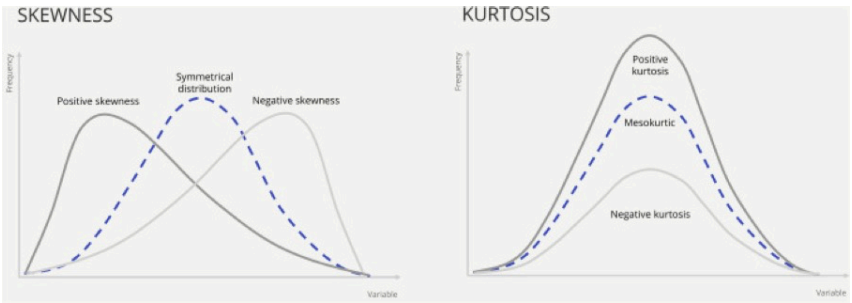


Figure 8: Skewness and kurtosis

Some units discussed in the lecture:

Example 4.5.
- regression metrics (MAE, MSE, RMSE, R2)
- classification metrics (accuracy, precision, recall, F1-score, sensitivity, specificity, ROC, AUC)
- Computer Vision metrics (PSNR, SSIM, IoU)
- timeseries related metrics (fit)

§4.1. Regression metrics

Mean absolute Error (MAE) is the average of the absolute differences between predicted and actual values.

Definition 4.1.1.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- y_i = measured value
- $\hat{y}_i$ = predicted value
- n = # of observations

Mean Squared Error (MSE) is the average of the squared differences between predicted and actual values.

Definition 4.1.2.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i = measured value
- $\hat{y}_i$ = predicted value
- n = # of observations

Root mean squared error (RMSE) is the square root of the mean of the squared differences between predicted and actual values.

Definition 4.1.3.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- y_i = measured value
- $\hat{y}_i$ = predicted value
- n = # of observations

R^2 or the coefficient of determination is the proportion of variance in the dependent variable that can be explained by the independent variable(s).

Definition 4.1.4.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} = 1 - \left(\frac{\mathbb{S}_{\text{res}}}{\mathbb{S}_{\text{tot}}}\right)$$

- y_i = measured value
- $\hat{y}_i$ = predicted value
- n = # of observations
- $\overline{y}_i$ = sample mean of the y_i's
- $\mathbb{S}_{\text{res}}$ = residual sum of squares
- $\mathbb{S}_{\text{tot}}$ = total sum of squares

Comparison of **MSE**, **MAE** and $\boldsymbol{R^2}$:

MSE	**MAE**	$\boldsymbol{R^2}$
based on square eror	based on the absolute value	based on correlation between actual and predicted value
value lies between 0 and $+\infty$	value lies between 0 and $+\infty$	value lies between $-\infty$ and 1
most sensitive to outliers	least sensitive to outliers	in-between sensitive to outliers
small value indicates better model	small value indicates better model	value near 1 indicates better model

§4.2. Classification metrics

When performing classification we have different outcomes, they can be seen in the Figure 9. An important definition is the types of errors that can occur:

Example 4.2.1.

- False positive (**FP**): incorrectly predicted positive class (**Type I error**)
- False negative (**FN**): incorrectly predicted negative class (**Type II error**)

		actual condition	
		true	false
test outcome	positive	true positive	false positive
	negative	false negative	true negative

Figure 9: False positive and false negative etc...

§4.3. Accuracy, recall, precision and F1-score

Accuracy is the ratio of correctly predicted instances to the total instances.

Definition 4.3.1 (Accuracy).

$$\text{Accuracy} = \frac{TP + TN}{\sum(TP + TN + FP + FN)}$$

Recall *(or sensitivity)* is the ratio of correctly predicted positive observations to all actual positive observations.

Definition 4.3.2 (Recall/sensitivity).

$$\text{Recall} = \frac{\sum TP}{\sum(TP + FN)}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Definition 4.3.3 (Precision).

$$\text{Precision} = \frac{\sum TP}{\sum(TP + FP)}$$

F1-score is the weighted average of precision and recall.

Definition 4.3.4 (F1-score).

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

When designing the classifier there are some rules of thumb to follow:

Example 4.3.1.

- **recall**: maximize this if false negatives are more costly
- **precision**: maximize this if false positives are more costly
- **F1 score**: maximize this if costs are similar and you want a balanced performance

We can visualize this with a ROC-curve[1]:

[1] ROC: receiver operating characteristic

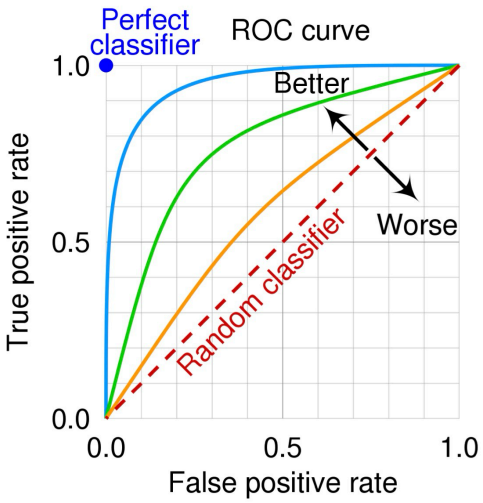


Figure 10: ROC-curve

Example 4.3.2.

- **High Sensitivity** & **Low Specificity**: you capture most positive instances but might misclassify many negative instances as positive. Appropriate only when detecting all positive instances is crucial, even if it raises false alarms
- **High Specificity** & **Low Sensitivity**: you correctly identify negative instances but might miss many positive instances. Appropriate when avoiding false positives is crucial, even if it means increasing the false negatives
- **High Sensitivity** & **High Specificity**: you rock!

§4.4. Metrics for computer vision

In computer vision we have some other metrics that are used. They are **PSNR**, **SSIM** and **IoU**.

Peak signal-to-noise ratio (PSNR) is a *maximum possible power of a signal / power of the noise that corrupts the signal.* The formula for PSNR is:

Definition 4.4.1 (PSNR).

$$\text{PSNR} = 10 * \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

- MAX = maximum possible pixel value of the image (usually 255 for 8-bit images)
- MSE = Mean Squared Error between the original and the compressed image

SSIM (Structural Similarity Index) is a *type of similarity between two images; intuitively, prediction of the quality of the second image as a distorted version of the first one, intended as distortion-free* The formula for SSIM is:

Definition 4.4.2 (SSIM).

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

- μ_x, μ_y = average pixel intensity of the two images
- σ_x^2, σ_y^2 = variance of the two images
- σ_{xy} = covariance between the two images
- c_1, c_2 = constant used to stabilize the division

Intersection over Union (IoU) is a metric used to evaluate the performance of an object detection model. It measures the overlap between the predicted bounding box and the ground truth bounding box. The formula for IoU is:

Definition 4.4.3 (IoU).

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{TP}{TP + FP + FN}$$

- Area of overlap = area of the intersection between the predicted and ground truth bounding boxes
- Area of union = area of the union between the predicted and ground truth bounding boxes

§4.5. Metrics for time series

In addition to MSE, RMSE and MAE we also have (goodness of) fit.

Definition 4.5.1 (Fit).

$$\text{fit} = 100 \cdot \left(1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}\right)$$

§5. Maximum likelihood (Lecture 5)

Why something more than least squares?

- (I.e., why do we introduce maximum likelihood estimators?)

answer: to include information about the statistical distribution of the measurement noise (and also to facilitate quantification of uncertainties)
caveat: by extending LS into ML, we "inject" assumptions. If the assumptions are good then ok, if they are bad then "not so ok"

§5.1. Types of probability distributions

Example 5.1.1.

- Uniform
- Exponential
- Poisson
- Gamma,
- Beta
- Binomial
- Log-normal
- Normal

§5.1.1. Info

Below you can see all the definitions of the distributions. See Definition 5.1.1.1– 5.1.1.6 And on page 13 they have all been visualized in the form of figures.

Definition 5.1.1.1 (Uniform).

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if a} \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Definition 5.1.1.2 (Exponential).

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

Definition 5.1.1.3 (Binomial).

$$f(x|n,p) = \frac{n!}{(n-x)!x!} p^{x(1-p)^{n-x}}$$

- n = number of trials
- x = number of successes
- p = probability of success for each trial

Definition 5.1.1.4 (Poisson).

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Definition 5.1.1.5 (Beta).

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \left(x^{\alpha-1}(1-x)^{\beta-1}\right)$$

- α and β = shape parameters
- B(·,·) = beta function

Definition 5.1.1.6 (Normal).

$$f(x|\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- mu = mean
- sigma = standard deviation

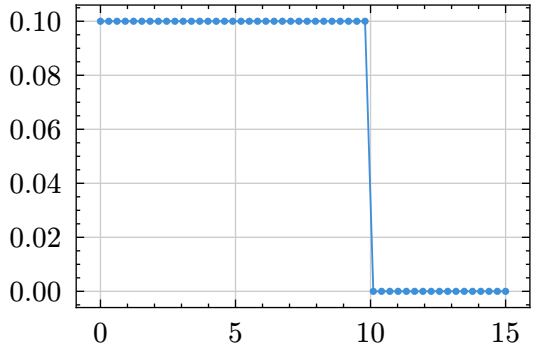


Figure 11: Uniform Distribution U(0, 10)

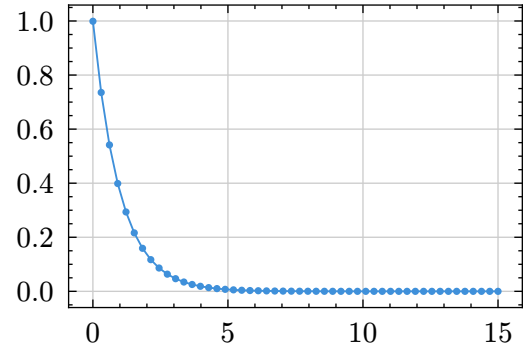


Figure 12: Exponential Distribution $\lambda = 1$

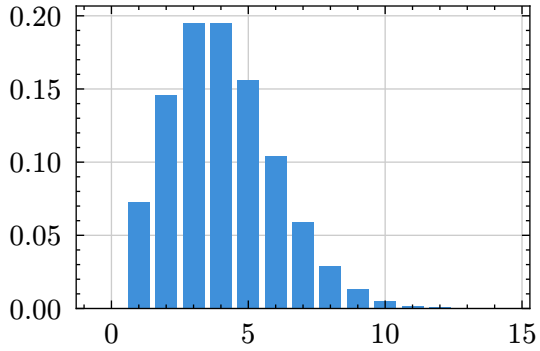


Figure 13: Poisson Distribution $\lambda \approx 4$

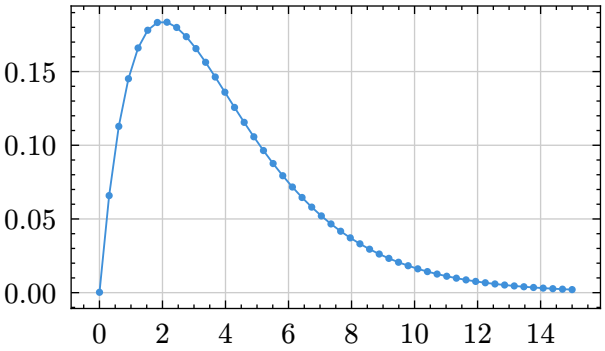


Figure 14: Gamma Distribution k=2, ϑ=2

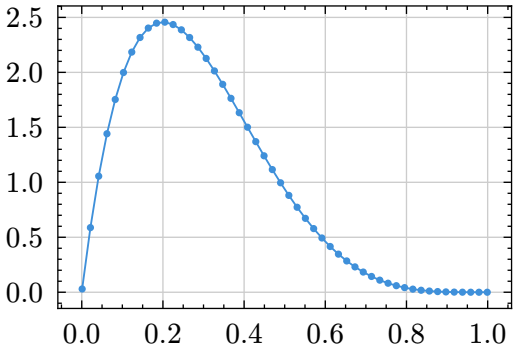


Figure 15: Beta Distribution α=2, β=5 (scaled)

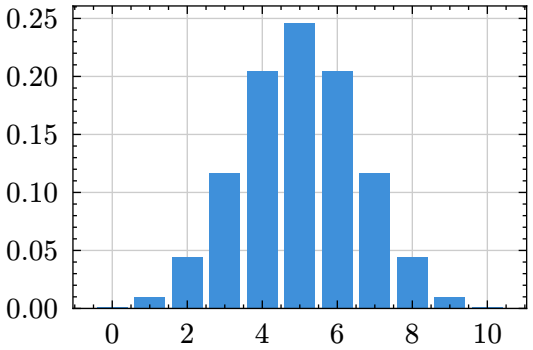


Figure 16: Binomial Distribution n=10, p=0.5

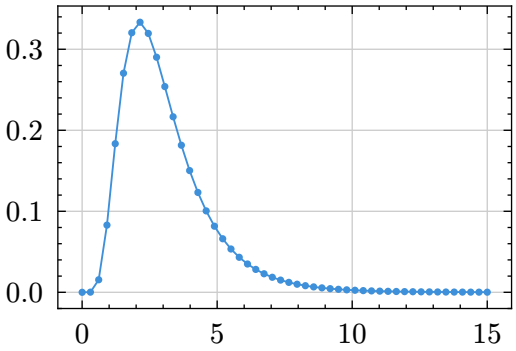


Figure 17: Log-normal Distribution μ=1, σ=0.5

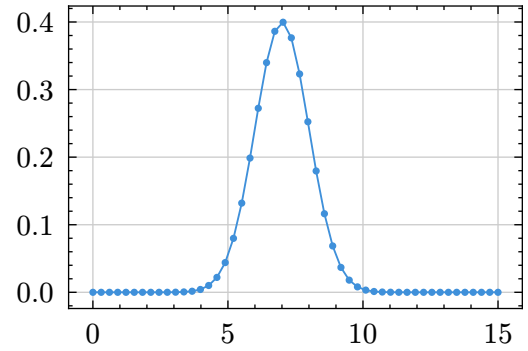


Figure 18: Normal Distribution μ=7, σ=1

§5.2. Probability vs. likelyhood

The basic notation for probability and likelyhood is given by:

Definition 5.2.1.
- P(data ; parameters) = probability (θ fixed, y to be generated)
- L(parameters ; data) = likelihood (θ to be estimated, y already collected)
  - **IMPORTANT**: P(data ; parameters) $\neq$ L(parameters ; data)

P vs. L in words:
They are different in how they treat data and parameters: probabilities describe the probability of data given fixed parameters, while likelihood functions de- scribe the likelihood of parameters given fixed data

Probability Function:
- assigns probabilities to outcomes of a random variable
- the parameters are thus fixed, the data is a variable *(thus this concept refers to before the data collection process)*
- integrates to 1

Likelihood Function:
- measures how well the parameters of the statistical model explains the observed data
- the data is thus fixed, the parameters is a variable *($\Rightarrow$ suitable for data analysis)*
- does not integrate to

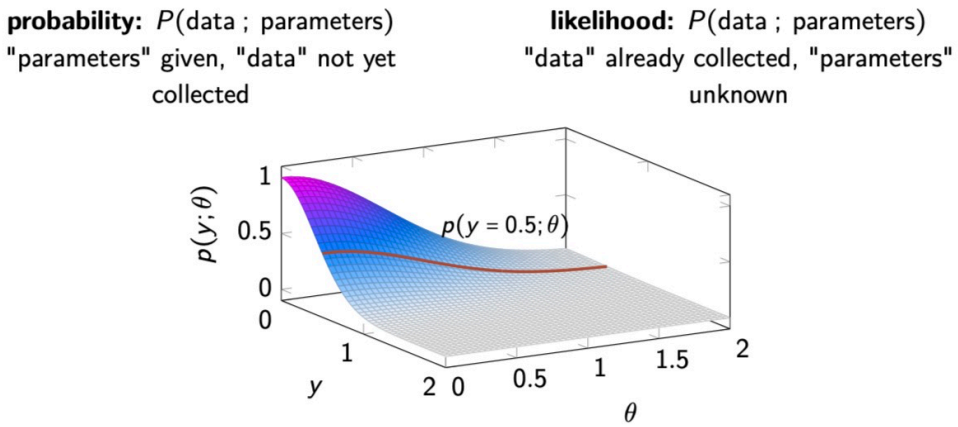


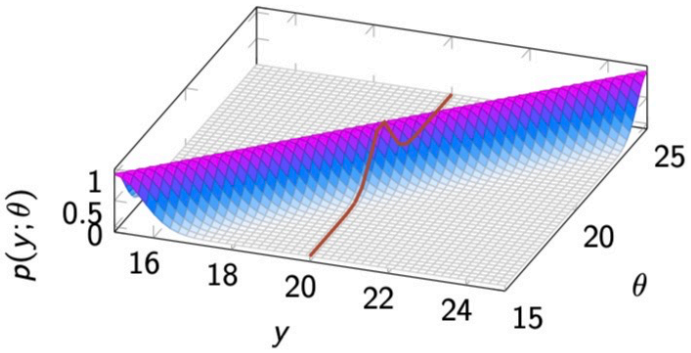
Figure 19: Probability vs. Likelihood

§5.3. Maximum likelihood

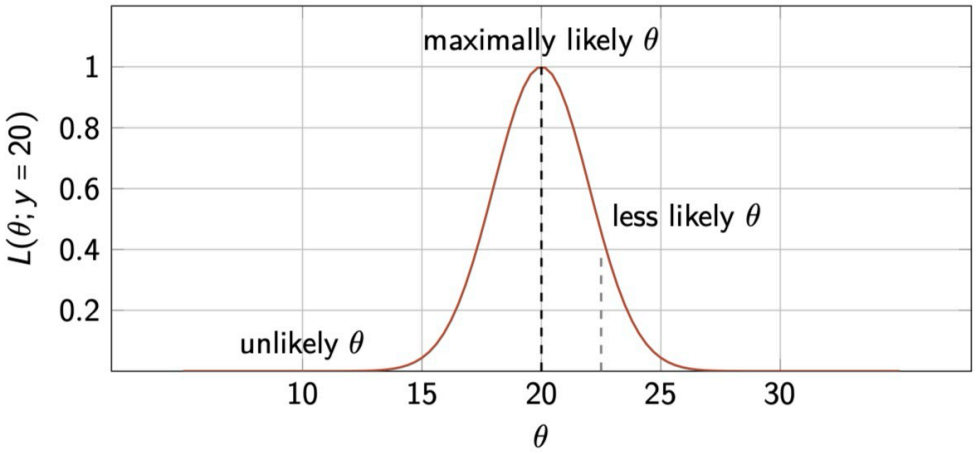
Towards Maximum Likelihood via an intuitive example

Scenario:

- want to estimate the unknown room temperature θ
- our temperature sensor measures $y = \theta + e$, with e additive Gaussian noise
- is it more likely that $\theta = 20$ or that $\theta = 10$?



Towards Maximum likelihood via another plot



§6. Maximum a posteriori (Lecture 6)

Maximum a posteriori comes from Bayes rule.

Definition 6.1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Then to get a posteriori we substitute A and B with more interesting names:

Definition 6.2.

- $A = \theta$
- $P(A) = P(\theta) =$ prior on θ
- $B = y$
- $P(B) = P(Y)$ evidence *(works as a normalization constant)*
- $P(B|A) = P(h|\theta) =$ likelihood

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \propto P(y|\theta)P(\theta)$$

§6.1. Maximum A posteriori = mode of the posterior distribution

Definition 6.1.1.

$\hat{\theta}_{\text{MAP}} := \arg\max_{\theta\in\Theta} \frac{P(y|\theta)P(\theta)}{P(y)} = \arg\max_{\theta\in\Theta} P(y|\theta)P(\theta)$

Example 6.1.1.

Suppose a medical test for a rare disease has a false positive rate of 5% and a false negative rate of 2%. The prevalence of the disease in the general population is 0.1%. If a person tests positive for the disease, what is the (posterior) probability that the person actually has the disease?

- Calculation: Let
  - A = Has disease
  - B = Positive result
  - $P(B|A) = 0.98$
  - $P(\neg B|\ \neg A) = 0.95$
  - $P(A) = 0.001$

This gives us:

$$P(B) = P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)$$

from:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Answer: $P(\theta|\neg\theta) = 1.9\%$

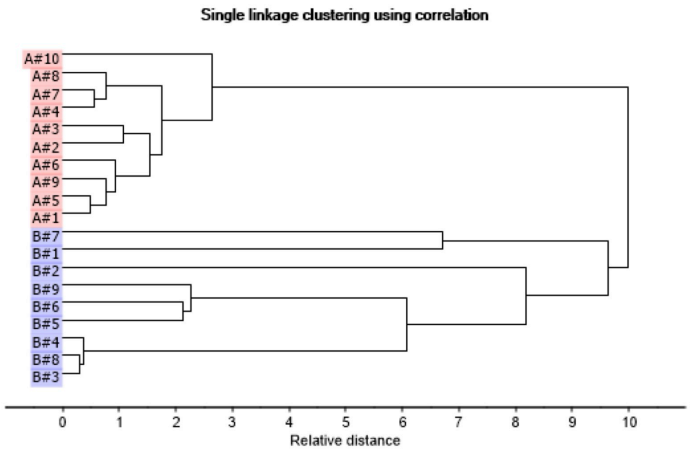
§6.2. Important disctinction to keep always in mind

- **LS** means "state the fitting problem as just a geometrical one"
- **ML** means "add the assumption that there is an underlying probability distribution of the data"
- **MAP** means "moreover add also a weighting based on an underlying prior assumption on the distribution of the parameters"

§7. How to visualize results (Lecture 7)

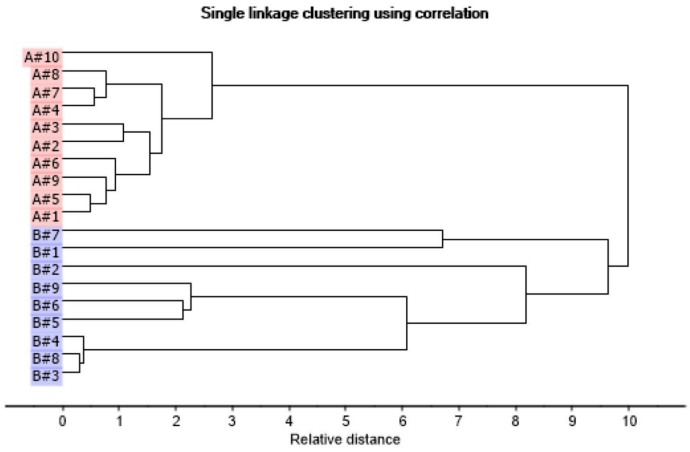
Case "not so many samples" and "single clusterer" $\Longrightarrow$ use a dendrogram

= a branching diagram that shows the hierarchical relationship between samples



Case "not so many samples" and "single clusterer" $\Longrightarrow$ use a dendrogram

= a branching diagram that shows the hierarchical relationship between samples



There are many more ways to visualize things but they are somewhat covered in the other parts by showing them, like residuals for regression algorithms and such.

An important part of visualization is to visualize **uncertainty**, this is done by showing **intervals** and **quantiles**.

§8. Basic statistics

Hypothesis testing is testing whether an observed difference is due to some effect or if it is due to random variation.

Definition 8.1 (Null hypothesis).
a statement that is assumed to be true unless there is strong evidence against it

Definition 8.2 (Alternative hypothesis).
what is accepted if the null hypothesis is rejected

Definition 8.3 (p-value).
p-value $= \mathbb{P}[X \geq x \mid H_0]$ = probability that things will be as extreme or worse than the given value x given the hypothesis H_0 and assuming the model $\mathrm{P}[\cdot]$

Definition 8.4 (p-value in words).
probability of obtaining test results
at least as extreme as the results actually observed
under the assumption of a probabilistic model
and that the null hypothesis is correct

Important messages about p-values:

- a conclusion does not become immediately true when the p-value passes the threshold
- be skeptical when you see p-values
- p-values rely on assumptions besides the tested hypothesis ⇒ do not forget keeping these assumptions in mind
- "absence of evidence is not evidence of absence"

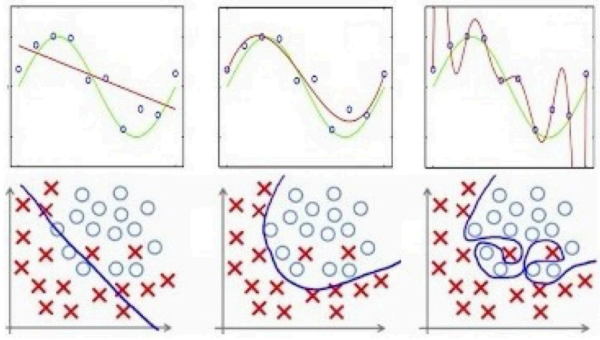
§8.1. UMP

There was a section about UMP testing, but I skipped it not bothering to learn everything in the course...

§9. The bias vs variance tradeoff

§9.1. Ockham's razor

Definition 9.1.1 (Occam's razor).
explaining a thing with no more assumptions should be made than are necessary



underfitting = a model that misses the fundamental features of the data
overfitting = a model that follows the data too much, and has thus too many features

ideal model: $y_t = f_0(u_t) + e_t$
our model: $y_t = f(u_t, \theta) + e_t$

Definition 9.1.2 (underfitting & overfitting).
- **underfitting**: a $\hat{\theta}$ that misses the fundamental features of f_0
- **overfitting**: a $\hat{\theta}$ that makes $\hat{f}$ follow e instead of f_0

Decomposing the MSE in two interesting terms

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{\theta}-\theta\|^2\right] &= \mathbb{E}\left[\|\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\|^2\right] \\
&= \mathbb{E}\left[\|\mathcal{V}+\mathcal{B}\|^2\right] && \begin{cases} \mathcal{V} := \widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right] \\ \mathcal{B} := \mathbb{E}\left[\widehat{\theta}\right]-\theta \end{cases} \\
&= \mathbb{E}\left[(\mathcal{V}+\mathcal{B})^T(\mathcal{V}+\mathcal{B})\right] \\
&= \mathbb{E}\left[\|\mathcal{V}\|^2+\|\mathcal{B}\|^2+2\mathcal{V}^T\mathcal{B}\right] && \mathbb{E}\left[\mathcal{V}^T\mathcal{B}\right] = \mathbf{0} \\
&= \mathbb{E}\left[\|\mathcal{V}\|^2\right]+\|\mathcal{B}\|^2 \\
&= \text{"variance"} + \text{"bias}^2\text{"}
\end{aligned}
$$

§10. More indept of PCA

This has been covered in great detail in the other courses so I will not bother with anything here...

§11. More PCA

This was only mathematical examples, boring... he he...

§12. Time series Modelling

First part was simply recap of ODE's.

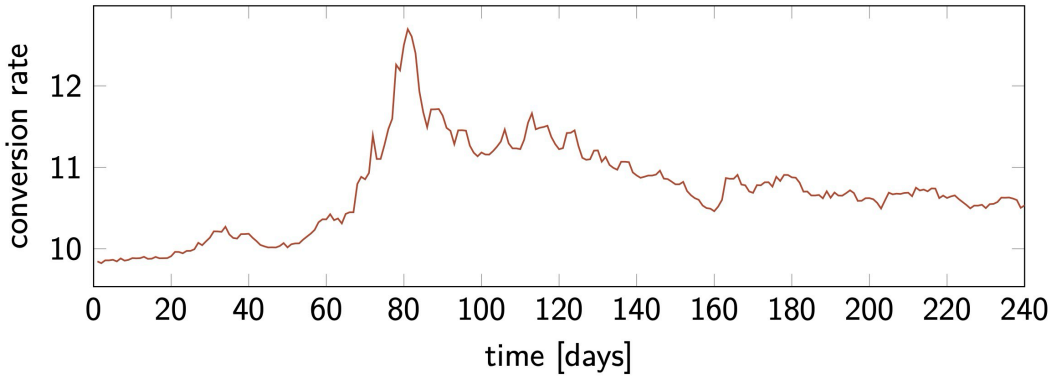
The important thing about time series modelling is for forecasting, meaning we want to predict future events. Here we focus on some LTI models:

- AR
- MA
- ARMA
- ARIMA

This chapter seemed to be a lot of repetition of Cybernetics introduction and Control Theory, I did not bother taking notes here...

§13. Time series classification

Most commonly used DMD-based features



- eigenvalues (real part = growth rates, imaginary part = oscillations)
- mode amplitudes
- reconstruction errors

§14. Validity of a model

Definition 14.1 (Internal validity).
does the model include all the important inputs that affect the output?

Definition 14.2 (External validity).
are the results generisable to other situations or groups?

Definition 14.3 (Construct validity).
does the model measure what it is supposed to measure?

Definition 14.4 (Population validity).
can the results be generalized to different group of people?

Example 14.1 (Partner 1).
I would like to make a potential partner fall in love with me, but I don't know how to do it. I then ask for advice from my three nephews, who are 4, 5, and 6 years old. They tell me they would fall in love with me if I buy them a bag of potato chips, sing a fairy tale, and dress as a bunny. I follow their advice when approaching my potential partner, but this makes her/him think I am a weirdo. Which type of validity was missing in my model of "how to approach people"?

Example 14.2 (Partner 2).
I would like to make a potential partner fall in love with me, but I don't know how to do it. I then ask for advice from my three nephews, who are 4, 5, and 6 years old. They tell me they would fall in love with me if I buy them a bag of potato chips, sing a fairy tale, and dress as a bunny. I follow their advice when approaching my potential partner, but this makes her/him think I am a weirdo. Which type of validity was missing in my model of "how to approach people"?

The rest of the chapter handled things regarding splitting in training test and validation data sets.

§15. Classification and discrimination, PCA and PLS-DA (Lecture 22)

§15.1. Cross model validation

- A common objective in multivariate modelling/machine learning/AI is to find the "best" model for a subset of features/variables
- Most methods have some built-in approach to avoid overfitting (cross-validation, regularization, tree-pruning,...)
- Nevertheless, when the same model is used for optimization of error and finding the best subset, this may lead to too optimistic results
- Another aspect is the false discovery rate in the case of many variables, e.g. 30000 genes for a limited number of patients
- One robust solution to these problems is Cross Model Validation (CMV), also called double cross validation

§15.2. Cross Model Validation (CMV)

- Take out a subset of the samples
- Cross validate the remaining
- Predict the samples kept out
- Repeat until all samples have been kept out
- Compare $RMSE_{\text{training}}$, $RMSE_{CV}$ and $RMSE_{CMV}$

CVM combined with jack-knifing is also very effective in screening out all non-relevant features as a first step in feature selection

§15.3. Jackknifing

- Jack-knifing uses the mean of all sub-models as basis for the variance estimates.
- Cross validation uses the model on all objects(more intuitive?)
- The difference between them is negligible in most practical applications

				Bootstrap						
	Segm 1	Segm 2	Segm 3	Segm 4	Segm 5	Segm 6	Segm 7	Segm 8	Segm 9	Segm 10
	2	5	8	8	2	6	3	5	3	7
	7	9	4	10	1	5	4	4	7	3
	4	9	9	6	9	6	8	9	4	9
	6	7	6	9	2	4	7	1	10	7
	2	9	4	2	3	5	5	8	8	2
	7	7	8	10	7	3	6	10	5	3
	4	4	6	3	3	6	8	10	8	7
	9	3	5	3	5	8	1	8	3	7
	9	4	7	9	1	6	7	5	5	4
	6	6	7	8	10	7	1	5	10	6
Unique objects	5	6	6	6	7	6	7	6	6	6
0.63 = theoretical										
				Jack-knife						
		1	1	1	1	1	1	1	1	1
	2		2	2	2	2	2	2	2	2
	3	3		3	3	3	3	3	3	3
	4	4	4		4	4	4	4	4	4
	5	5	5	5		5	5	5	5	5
	6	6	6	6	6		6	6	6	6
	7	7	7	7	7	7		7	7	7
	8	8	8	8	8	8	8		8	8
	9	9	9	9	9	9	9	9		9
	10	10	10	10	10	10	10	10	10	

Figure 25: Jackknifing

§15.4. Classification and discrimination

- **clustering**: deciding how to group some object together
- **classification**: deciding what kind of object something is
- **discrimination**: Measures how well the model separates data from different classes

Note: often "discrimination" is considered as a specific "classification"

§15.5. The two overarching strategies

- build local models of the various classes
- build models of the boundaries between the various classes

§15.6. Strategy 1: build local classes

Typical workflow:

- Build a model for each class
- Samples inside the critical limits are accepted
- Others are rejected not classified
- Methods: PCA, ICA, CLustering, one class SVM, autoencoders

Drawbacks: the individual models don't know about the other classes and there is no objective function to discriminate between them in the modelling phase

§15.7. Strategy 2: create objective functions to discriminate between classes

Frequently applied methods:

- PLS-DA

- LDA (Linear Discriminant Analysis)
- Logistic regression
- SVM (Support Vector Machines)
- Clustering methods ()
- ANN/CNN (Artificial Neural Networks/Convolutional Neural Networks)
- Classification trees/Random Forest/XGBoost ...

§15.8. Disclaimer

There is more in this lecture but I will get back to it if time. Now I have to get to the other lectures to skip the least amount of materisl.

§16. Support Vector machines (SVM) (Lecture 23)

Used to classify (and also regression) that originates from machine learning. It uses a **kernel function** to map from originnal space to a feature space where a hyperplane can divide the classes. It is suited for non-linear problems and non-homogenous classes. However SVM only takes into account the samples at the boundaries (called support vectors) to establish the rule for discrimination.

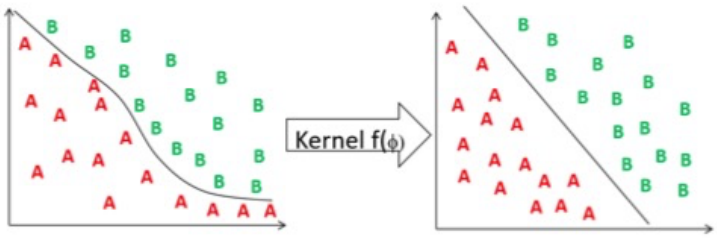
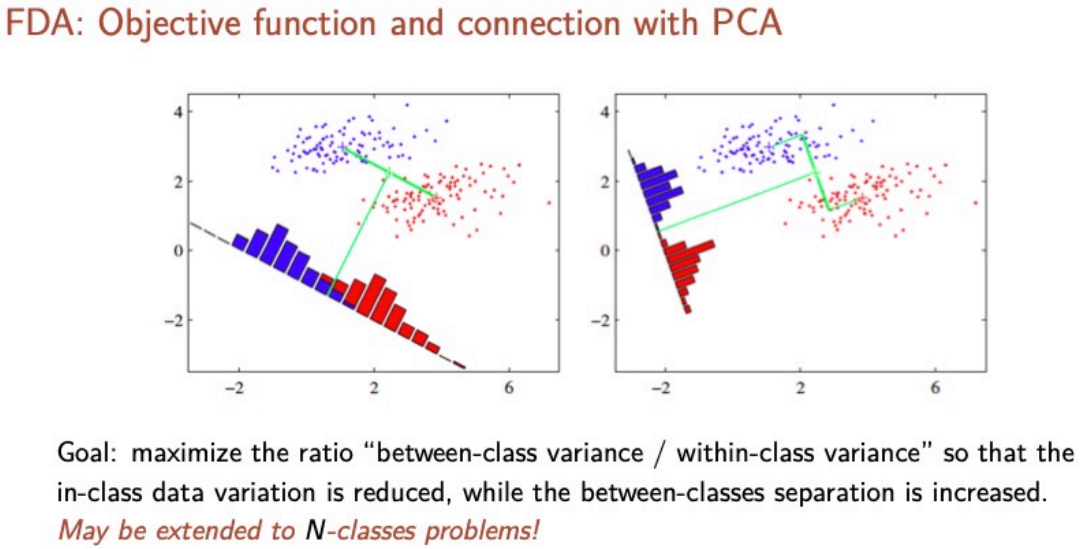


Figure 26: Kernel function

§16.1. Linear and Fisher Discriminant Analyses (LDA and FDA)

- **FDA**: which linear combination of features **w'x** best separates the data?
- **LDA**: same thing, but motivated by some specific statistical assumptions on the distribution of the data.



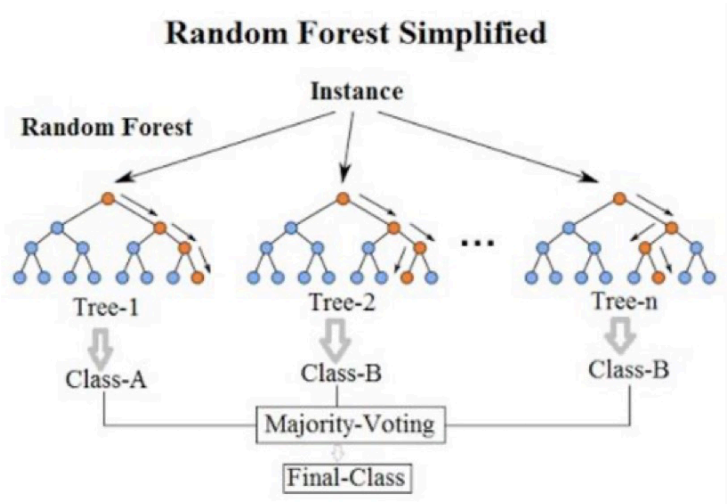


Figure 28: Random forest

§16.2. Tree based classifictaion models

§16.2.1. Decision trees

This has already been covered by the course methods in artificial intelligence and chemometrics. Details are provided there.

§16.2.2. Random forests

Random forests are an ensemble of decision trees. The idea is to combine the predictions of multiple trees to improve the overall performance. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees.

Properties of **random forest**:

- It is one of the best methods in accuracy
- It runs efficiently on large data bases
- It can handle thousands of input variables without variable deletion
- It gives estimates of what variables are important in the classification
- It generates an internal unbiased estimate of the generalization error as the forest building progresses
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing
- It has methods for balancing error in class population unbalanced data sets

§17. Singular Values Decomposition (Lecture 24)

This lecture seemed like alot of algebra, not sure what to take from it. Might not be important, will read up on it if I get time before the exam.

§18. ANN/CNN (Lecture 25)

Artificial Nerual Networks (ANN) and Convolutional Neural Networks (CNN) are two types of neural networks used in machine learning and deep learning. They are designed to recognize patterns and make predictions based on input data.

Universal approximation theorem: A feed forward neural network with a linear output layer and at least one hidden layer with any "squashing" activation function can approximate any continuous function.
However, no guarantee on the learnability of the function primarily due to:

- finding global minima
- due to over fitting

Several activaiton functions can be used, but the most common one is the sigmoid function, which is defined as:

Definition 18.1 (Sigmoid function).

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Convolution network use a convolution between some sort of "mask" going over the "image". This results in reducing the size of the network, show in Figure 29. Maxpooling is kind of the same but instead of convolution it simply takes the max of the given grid size e.g. 2 as shown in Figure 30.

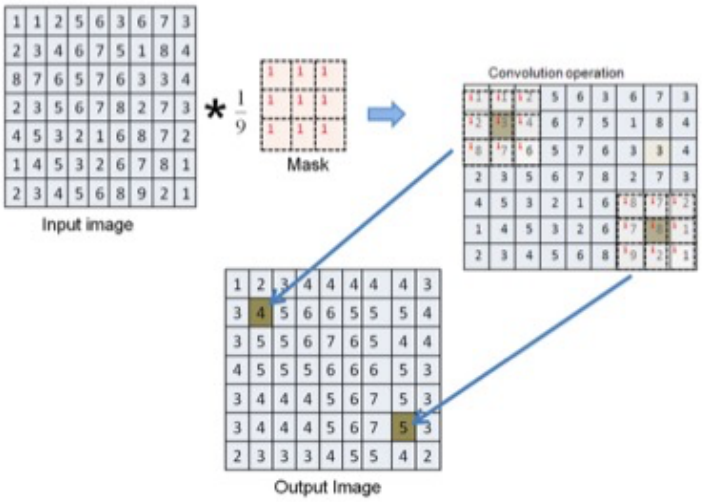


Figure 29: Convolutional network

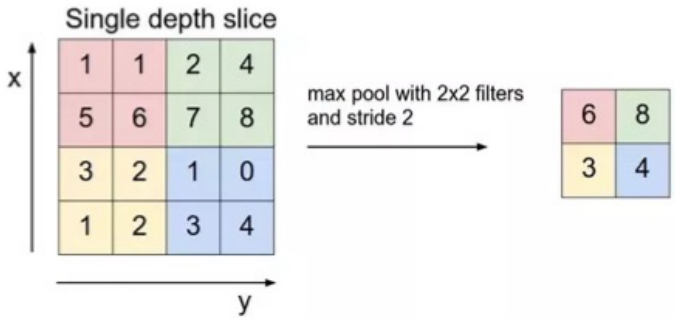


Figure 30: Maxpooling

We can also see an example structure of a CNN in Figure 31.

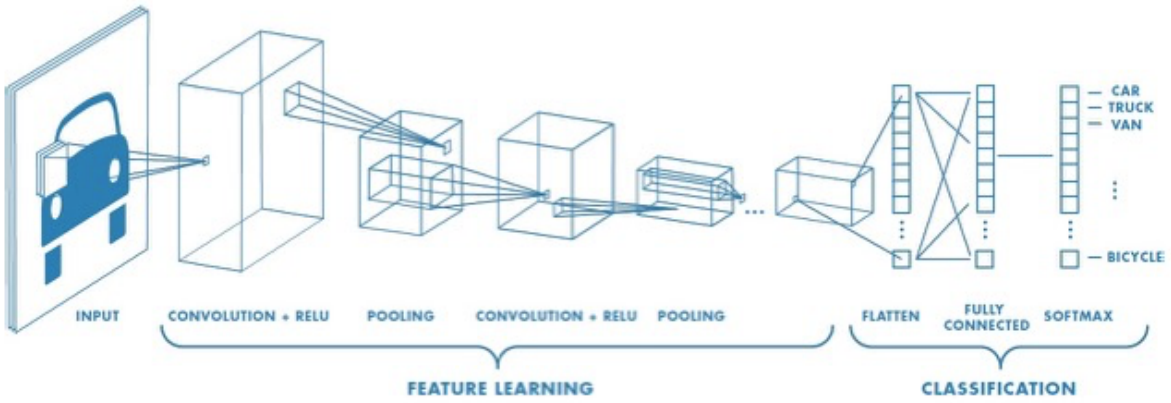


Figure 31: CNN structure

§18.1. Style transfer

Style transfer is a technique used in deep learning to apply the style of one image to the content of another image. It involves using convolutional neural networks (CNNs) to extract features from both the content and style images, and then combining these features to create a new image that retains the content of the original image while adopting the style of the style image.

§18.2. Autoencoders

Autoencoders are a type of neural network used for unsupervised learning. They are designed to learn efficient representations of data by compressing it into a lower-dimensional space and then reconstructing it back to the original space. The main components of an autoencoder are the encoder, which compresses the input data, and the decoder, which reconstructs the data from the compressed representation.

§19. Potential exam questions *(provided on BlackBoard)*

Q&A 19.1.

- *What do we mean with internal and external validity? May you make some practical examples?*
  - Internal validity refers to the degree to which a study accurately identifies a causal relationship between variables, by controlling for confounders. For example, a lab experiment with randomized groups. External validity refers to the generalizability of findings to other populations or settings, e.g., applying lab results to real-world patients.

Q&A 19.2.

- *What are least squares estimators? What is the geometrical intuition behind them? And how can they be formulated mathematically?*

- Least squares estimators minimize the sum of squared differences between observed and predicted values. Geometrically, this corresponds to projecting the observation vector onto the space spanned by the regressors. Mathematically: $\hat{\vartheta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$.

Q&A 19.3.

- *What are maximum likelihood estimators? What is the probabilistic intuition behind them? And how can they be formulated mathematically?*
  - Maximum likelihood estimators choose parameters that make the observed data most probable. The idea is to find the parameter ϑ that maximizes the likelihood function $L(\vartheta|\text{data})$. For independent samples, this often involves maximizing the product (or log-sum) of densities.

Q&A 19.4.

- *Why should one divide a dataset into training, test, and validation sets? And how should one select such sets from the original dataset?*
  - Splitting data ensures that the model is trained, tuned, and tested on separate sets to avoid overfitting and evaluate generalization. Randomly shuffle and divide data, or use stratified sampling if class imbalance exists.

Q&A 19.5.

- *What does cross validation mean? How should it be used? And why should one use a cross validation approach instead of using a training, test, and validation sets based approach?*
  - Cross-validation involves dividing data into k folds, training on k-1 and testing on the remaining one, iteratively. It gives a robust estimate of model performance and is preferred when limited data makes a fixed split unreliable.

Q&A 19.6.

- *What are maximum a posteriori estimators? What is the probabilistic intuition behind them? And how can they be formulated mathematically?*
  - MAP estimators maximize the posterior probability of parameters, combining prior beliefs and observed data. It's like MLE but includes a prior: $\hat{\vartheta} = \text{argmax } P(\vartheta|\text{data}) \propto P(\text{data}|\vartheta)P(\vartheta)$.

Q&A 19.7.

- *What are the conditions about existence and uniqueness of the LS, ML and MAP estimates?*
  - LS and ML estimates require full column rank in the design matrix for a unique solution. MAP adds a prior, which can ensure uniqueness even in ill-posed problems if the prior is well-formed.

Q&A 19.8.
- *Derive the maximum likelihood estimator for the separable problem* $yi = \vartheta ui + vi$ *with* $vi \sim N(0, \sigma^2)$ *with* σ^2 *known,* ϑ *unknown and deterministic*
  - The likelihood is a product of Gaussians. Taking the log, differentiating, and solving gives $\hat{\vartheta} = (\Sigma ui^2)^{-1}\Sigma uiyi$, which minimizes squared residuals, same as LS here.

Q&A 19.9.
- *Derive and comment the bias-variance tradeoff*
  - Expected error = bias^2 + variance + irreducible noise. Complex models have low bias, high variance (overfit); simple models have high bias, low variance (underfit). The goal is a balanced model minimizing total error.

Q&A 19.10.
- *What do we expect to see when training and testing different estimators with different model order complexities? How should we account for the effects that we see on the statistical performance indexes?*
  - As complexity increases, training error typically decreases, but test error may rise due to overfitting. This is seen as a U-shaped curve in validation error. Use cross-validation and penalized metrics (e.g., AIC, BIC) to choose optimal complexity.

Q&A 19.11.
- *Which statistical performance indexes would you consider when dealing with a regression problem? And which peculiarities / usages do they have?*
  - MSE and RMSE measure average error; MAE is less sensitive to outliers; R^2 indicates explained variance. Choose based on robustness or interpretability needs.

Q&A 19.12.
- *Which statistical performance indexes would you consider when dealing with a classification problem? And which peculiarities / usages do they have?*
  - Accuracy, precision, recall, F1 score. Precision-recall better for imbalanced data. ROC AUC shows tradeoff over thresholds.

Q&A 19.13.
- *What does "design of experiments" mean? And "factorial design"? Which alternative factorial-design based alternatives do you know, and what are the tradeoffs among them?*
  - DOE is planning experiments to extract information efficiently. Full factorial tests all variable combinations. Alternatives like fractional factorial reduce experiments at cost of detail.

Q&A 19.14.
- *What does PCA mean from a geometrical point of view? How is it formulated mathematically, and how does it connect with SVD?*
  - PCA finds directions (principal components) of maximal variance. Mathematically via eigendecomposition or SVD of the data matrix. PCs are orthogonal axes of variation.

Q&A 19.15.
- *What are the uses of PCA? And how can its results be interpreted?*
  - PCA is used for dimensionality reduction, denoising, and visualization. Loadings show how variables contribute; scores show sample projections.

Q&A 19.16.
- *What are the uses of the loadings plots and scores plots, in a PCA?*
  - Loadings plot shows how variables contribute to PCs. Scores plot shows patterns or clusters among samples in the reduced space.

Q&A 19.17.
- *How can one decide how many components should be used when analysing some data through PCA? And how can one decide whether a sample is an outlier or not, through PCA?*
  - Use scree plots or cumulative explained variance. Outliers are samples with high score distance or reconstruction error.

Q&A 19.18.
- *How does the Ockham's razor principle connect with the model order selection problem? Which alternative strategies can be used to solve the model order selection problem?*
  - Prefer the simplest model that explains the data well. Strategies include AIC, BIC, cross-validation, and hypothesis testing.

Q&A 19.19.
- *What does "rotated PCA" mean? How does this concept connect with PCA, from both geometrical and mathematical points of view?*
  - Rotation (e.g., varimax) simplifies interpretation by redistributing variance across PCs while keeping orthogonality. Enhances variable grouping.

Q&A 19.20.
- *Which type of problems does the ICA algorithm solve? Which assumptions does it require? And how does it work, from intuitive perspectives?*

- ICA separates mixed signals into independent sources (e.g., blind source separation). Assumes statistical independence and non-Gaussianity.

Q&A 19.21.
- *What does "total least squares" mean? How does this concept connect with least squares, from both geometrical and mathematical points of view?*
  - TLS accounts for noise in both X and y. Geometrically, it minimizes orthogonal distances to the model plane, unlike LS which only minimizes vertical distance.

Q&A 19.22.
- *What does ANOVA mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - ANOVA tests differences between group means. It partitions variance into within- and between-group components and uses F-tests for inference.

Q&A 19.23.
- *What does PLS mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - PLS projects X and y to latent variables that maximize covariance. Useful when predictors are many and collinear. Combines features of PCA and regression.

Q&A 19.24.
- *What does MLR mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - MLR models the linear relation between multiple inputs and an output. $y = X\beta + \varepsilon$, estimated by LS. Used for prediction and inference.

Q&A 19.25.
- *What does PCR mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - PCR uses PCA to reduce dimensionality of X, then applies regression on components. Helps when predictors are correlated.

Q&A 19.26.
- *How do PLS, MLR, and PCR relate to each other? In which cases does one expect one of them to work better than the other ones, and viceversa?*
  - MLR works if predictors are uncorrelated. PCR reduces dimensions blindly. PLS considers both X and y and often performs best with collinearity or few samples.

Q&A 19.27.
- *What is the NIPALS algorithm? How does it work, from a graphical perspective? Which advantages does it bring over SVD, when used to compute a PCA?*
  - NIPALS computes components iteratively, one at a time. Useful for large or incomplete datasets. More flexible than full SVD.

Q&A 19.28.
- *What does metamodelling mean? When would one want to use a metamodelling approach? What are the potential shortcomings of a metamodel?*
  - Metamodels approximate complex models using simpler ones. Useful for simulation or optimization. Downsides: potential inaccuracy, limited extrapolation.

Q&A 19.29.
- *What do "stationarity" and "ergodicity" mean? Why are these two concepts important when dealing with statistical analyses of time series? And what would the lack of stationarity and ergodicity imply in practice?*
  - Stationarity: statistical properties don't change over time. Ergodicity: averages over time ≈ averages over ensemble. Lack of these breaks model assumptions.

Q&A 19.30.
- *Which LTI model structures do you know that are suitable to do control-oriented modelling of discrete time MISO systems?*
  - ARX, ARMAX, OE, and Box-Jenkins. Choice depends on noise and model goals. ARX is simplest, BJ most flexible.

Q&A 19.31.
- *What is the principle behind prediction error methods? When should the focus be on prediction errors, when identifying a dynamical model?*
  - PEM minimizes the difference between observed and predicted outputs. Focus on prediction errors when the model is used for forecasting or simulation.

Q&A 19.32.
- *What are the implications of choosing an ARX, instead of an ARMAX, instead of an OE model structure when doing system identification? And what are the implications of choosing different model orders? How should one choose a specific structure and order?*
  - ARX is fast but sensitive to noise. ARMAX models noise more accurately. OE assumes noise-free input. Choose structure by validation and orders via AIC/BIC.

Q&A 19.33.

- *What are Hammerstein Wiener models, and what are their usages?*
  - They model nonlinear dynamics via static nonlinearity + linear dynamic blocks. Useful for control systems with known nonlinearity.

Q&A 19.34.
- *What is a p value? How should it be computed? What is its usage for? And its drawbacks?*
  - p-value is the probability of seeing data as extreme as observed, under H_0. Computed from test statistic distribution. Misuse includes binary cutoff thinking.

Q&A 19.35.
- *What is a statistical test? How can it be interpreted from geometrical perspectives? And from mathematical perspectives?*
  - A statistical test evaluates hypotheses using data. Geometrically, compares projections. Mathematically, tests whether a statistic lies in a rejection region.

Q&A 19.36.
- *What are the statistical performance indexes associated to a statistical test? And which concepts may one use to say that a test is "better" than another one?*
  - Power, significance level, Type I/II error rates. Better tests have higher power for fixed significance. UMP tests are optimal under specific conditions.

Q&A 19.37.
- *What are the differences between simple and composite hypotheses? How do the formulations of hypothesis testing algorithms change, depending on which type of hypothesis is considered?*
  - Simple: fully specified distributions. Composite: parameters unknown. Composite tests often require approximations or likelihood ratios.

Q&A 19.38.
- *What does the Linear Discriminant Analysis algorithm do? How? Which advantages and disadvantages does it have?*
  - LDA finds projection that separates classes based on class means and shared covariance. Works well with Gaussian data, struggles with nonlinear boundaries.

Q&A 19.39.
- *What does the Partial least squares discriminant analysis algorithm do? How? Which advantages and disadvantages does it have?*
  - PLS-DA adapts PLS for classification, modeling latent structures in predictors to distinguish classes. Good with multicollinearity, less interpretable.

Q&A 19.40.
- *What does the logistic regression algorithm do? How? Which advantages and disadvantages does it have?*
  - Models class probabilities using a sigmoid function over linear combinations. Interpretable but limited to linear decision boundaries.

Q&A 19.41.
- *What does the Support Vector Classification algorithm do? How? Which advantages and disadvantages does it have?*
  - SVC finds hyperplanes maximizing class margins. Effective in high dimensions. Requires tuning and can be slow for large datasets.

Q&A 19.42.
- *What is the kernel trick? Where may one use it, and why?*
  - Maps data to high dimensions via kernels without explicit computation. Enables nonlinear classification or PCA efficiently.

Q&A 19.43.
- *What does the K-means algorithm do? How? Which advantages and disadvantages does it have?*
  - K-means partitions data into k clusters by minimizing within-cluster variance. Simple and fast, but assumes spherical clusters and needs k known.

Q&A 19.44.
- *What does the DBSCAN algorithm do? How? Which advantages and disadvantages does it have? What are its differences with the k-means algorithm?*
  - Density-based clustering; identifies noise and arbitrarily shaped clusters. Unlike k-means, doesn't require number of clusters but sensitive to parameters.

Q&A 19.45.
- *What do decision trees and random forests do? How? Which advantages and disadvantages do they have?*
  - Trees split data based on features. Forests average over many trees to reduce overfitting. Trees are interpretable, forests are more accurate.

Q&A 19.46.
- *What is the need for features selection? And what are the differences between feature engineering and selection?*

- Reduces overfitting and improves interpretability. Selection chooses from existing features; engineering creates new informative features.

Q&A 19.47.

- *What are the pros and cons of using a wrapper method for feature selection, or an embedded method for the same sake?*
  - Wrappers use model feedback, often more accurate but slower. Embedded methods (e.g., Lasso) are faster but model-dependent.

Q&A 19.48.

- *Why is categorical cross entropy loss used instead of mean squared error as a cost function in classification problems as against the regression problems?*
  - Cross-entropy measures divergence between predicted and true class probabilities, offering better gradients and performance than MSE in classification.

Q&A 19.49.

- *Which regularization technique is more effective in feature selection? And why?*
  - L1 (Lasso) is better than L2 (Ridge) for selection because it shrinks some coefficients to zero, effectively removing features.