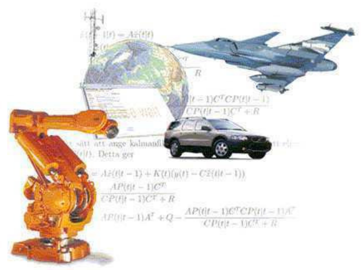
Welcome to Machine Learning 2013!!



Thomas Schön

Division of Automatic Control
Linköping University
Linköping, Sweden.

Email: schon@isy.liu.se,
Phone: 013 - 281373,
Office: House B, Entrance 27.

What is machine learning all about?

"Machine learning is about learning, reasoning and acting based on data."

Outline lecture 1

1. Introduction and some motivating examples
2. Course administration
3. Probability distributions and some basic ideas
    1. Exponential family
    2. Properties of the multivariate Gaussian
    3. Maximum Likelihood (ML) estimation
    4. Bayesian modeling
    5. Robust statistics ("heavy tails")
    6. Mixture of Gaussians

Problem classes

- **Supervised learning.** The training data consists of both input and output (target) data.
    - *Classification:* Discrete output variables.
    - *Regression:* Continuous output variables.
- **Unsupervised learning.** The training data consists of input data only.
    - *Clustering:* Discover groups of similar examples in data.
- **Reinforcement learning.** Finding suitable actions (control signals) in a given situation in order to maximize a reward. Close to control theory.

This course is focused on supervised learning.

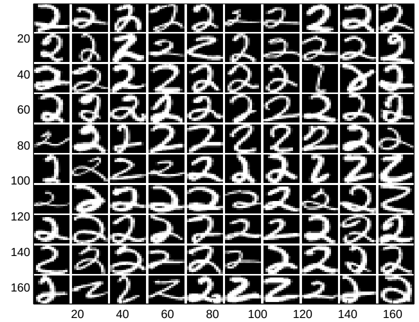
Example 1 – autonomous helicopter aerobatics

- Learning good controllers for tasks demonstrated by a human expert. Currently a hot topic in many areas (related to ILC).
- Includes learning a model, estimating the states, learning a controller

Pieter Abbeel, Adam Coates and Andrew Y. Ng. **Autonomous helicopter aerobatics through apprenticeship learning**, *International Journal of Robotics Research (IJRR)*, 29(13):1608-1639, November 2010.

Example 2 – handwritten digit classification

- Input data: 16×16 grayscale images.
- Task: classify each input image as accurately as possible.
- This data set will be used throughout the course.
- Solutions and their performance are summarized on `yann.lecun.com/exdb/mnist/`



Data set available from

`www-stat.stanford.edu/~tibs/ElemStatLearn/`

Example 3 – BNP for dynamical systems

BNP (lecture 11) offers flexible models capable of dealing with

- How many states should be used?
- How many modes? (i.e., hybrid systems)
- What if new modes/states arise over time?

E.B. Fox, E.B. Sudderth, M.I. Jordan, A.S. Willsky. **Sharing Features among Dynamical Systems with Beta Processes**, *Proceeding of Neural Information Processing Systems (NIPS)*, Vancouver, Canada December 2009.

Example 4 – animal detection and tracking (I/II)

Volvo förnyar system som ser faror i mörker

Volvo utvecklar sitt system för upptäckt av gående som riskerar att bli påkörda till att också fungera i mörker och för stora djur.

Jag möter Volvos tekniker i den tidiga skymningen i ett hägn med ett 30-tal älgar, dov- och kronhjortar. I en Volvo V70 har det monterats en infraröd kamera framför innerbackspegeln. Kameran är kopplad till en persondator inuti bilen.

Under körning längs vägen i hägnet läser datorn i bilen in de olika djuren från olika vinklar. Både i rörelse och stillastående.

Det är det första fältprovet i en utveckling som ska leda till att bilen själv lär sig känna igen att fara för att viltkollision hotar. Den ska då med ljud- och ljussignaler varna föraren. Om denne inte ingriper är det tänkt att bilen bromsar för att undanröja faran.

När ska systemet ingripa för djur?

– Där det finns risk för att förare och passagerare skadas, svarar Andreas Eidehall, som är säkerhetsexpert med djurdetektion som specialitet hos Volvo Personvagnar.

Han förklarar att svårast är att få systemet att inte ingripa i onödan. Därför vägs radarns (monterad i fronten) och den infraröda kamerans information samman.

Andreas Eidehall

– Systemet, som ännu inte har fått något namn, är tänkt att varna förare som är distraherade och ouppmärksamma.

ANDREAS EIDEHALL säger att nackdelen med de nattsynsystem som redan finns på marknaden är att de tvingar föraren att splittra sin uppmärksamhet mellan vindrutan och bildskärmen.

Det blir allt skummare i hägnet i takt med att solen sjunker i väst. Och allt svårare att uppfatta djuren som lockas till vägen med utlagt foder.

Andreas Eidehall berättar att viltolyckor är ett stort problem inte bara i Sverige, utan i många länder där det säljs Volvobilar. I vårt eget land skedde i fjol 40 000 viltolyckor, berättar han. Av dessa olyckor var 13 procent med älg.

”Systemet, som ännu inte har fått något namn, är tänkt att varna förare som är distraherade och ouppmärksamma.

Andreas Eidehall, säkerhetsexpert

När klockan har blivit 22.30 är det svårt att urskilja djuren som helt oblygt beträder vägen fastän bilen kommer rullande med påslaget halvljus.

JAG ÅKER MED i bilen och ser på datorns skärm hur väl det infraröda ljuset på gott håll avslöjar djuren i mörkret fastän jag knappt kan skönja dessa när jag tittar ut genom vindrutan.

– Vi ska träna systemet att känna igen djuren, säger Andreas Eidehall som tror att den kommersiella lanseringen i Volvos personbilar ligger några år bort.

JACQUES WALLNER
jacques.wallner@dn.se 08-738 19 61

GUIDE

Bilens dator läser in djuren från olika vinklar.

Automatskydd mot påkörning

- **Dagens system** ”pedestrian detection with full auto brake” känner igen gående och cyklande i dagsljus, varnar föraren om kollision hotar och ingriper med fullbroms om det behövs för att undvika en olycka.
- **Nästa steg** är att också upptäcka stora vilda djur på och längs vägen
- **Med kombinationen** radar och infraröd kamera upptäcks både människor och stora djur också när det är mörkt.
- **Tanken är att** täcka en 45 grader stor sektor framför bilen, räckvidden bör vara minst 75 meter.

Example 4 - animal detection and tracking (II/II)

Volvos automatiska skydd mot påkörning är under utveckling. Snart ska det kunna upptäcka stora djur på vägen. FOTO: ANDERS WEJROT

- Learning detectors for animals. boosting (lecture 8) promising technology for this.
- Sensor fusion between radar and infrared camera.

Field of machine learning

Top 3 conferences on general machine learning

1. Neural Information Processing Systems (NIPS)
2. International Conference on Machine Learning (ICML)
3. European Conference on Machine Learning (ECML) and Inter. Conf. on Artificial Intelligence and Statistics (AISTATS)

Top 3 journals on general machine learning

1. Journal of Machine Learning Research (JMLR)
2. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)
3. IEEE Trans. on Neural Networks (TNN)

For new (and non-peer reviewed) material see arXiv.org
`arxiv.org/list/stat.ML/recent`

Course administration

- Lecturers: Thomas Schön and Fredrik Lindsten
- Examiner: Thomas Schön
- 11 lectures (do not cover everything)
- We will try to provide examples of active research throughout the lectures (especially connections to "our" areas)
- Suggested exercises are provided for each lecture
- Written exam, 3 days (72 hours). Code of honor applies as usual
- All course information, including lecture material is available from the course home page
  `www.control.isy.liu.se/student/graduate/MachineLearning/`

Course administration - projects (3 hp)

- Voluntary and **must be based on a data set**.
- Project ideas: discuss with me for ideas or even better, make up your own!!
- Form teams (2-3 students/project).
- Project time line:

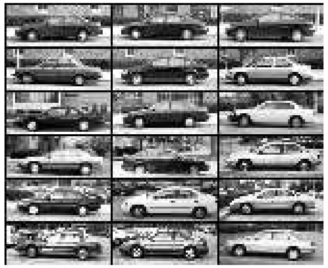
Date	Action
Mar. 20	Project proposals are due
Mar. 22	Project proposal presentation
Apr. 19	Final reports are due
Apr. 24	Final project presentations

- See course home page for details.
- Note that the deadline for NIPS is in the beginning of June.

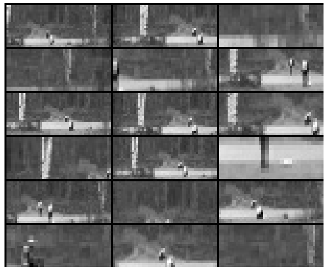
Project example from the 2011 edition

Detection and classification of cars in video images

Task: Train a detector/classifier, which can be used to detect, track and eventually classify different vehicles in the video recordings.



Positive training samples.



Negative training samples.

A semi-supervised tracker was also developed (see movie).

Wahlström, N. and Granström, K. **Detection and classification of cars in video images**, *Project report*, May, 2011.

Project example from the dynamic vision course

Helicopter pose estimation using a map

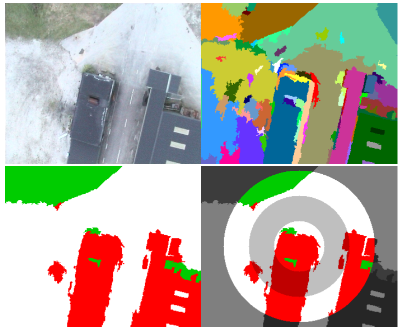


Image from on-board camera (top left), extracted superpixels (top right), superpixels classified as grass, asphalt or house (bottom left) and three circular regions used for computing the class histograms (bottom right).

Map over the operational area (top), manually classified reference map (bottom).

Fredrik Lindsten, Jonas Callmer, Henrik Ohlsson, David Törnqvist, Thomas B. Schön, Fredrik Gustafsson. **Geo-referencing for UAV Navigation using Environmental Classification**. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Anchorage, Alaska, USA, May 2010.

Course overview – Topics

1. Linear regression
2. Linear classification
3. Expectation Maximization (EM)
4. Neural networks
5. Gaussian processes (first BNP)
6. Support vector machines
7. Clustering
8. Approximate inference
9. Boosting
10. Graphical models
11. MCMC and sampling methods
12. Bayesian nonparametrics (BNP)

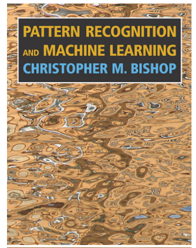
Literature

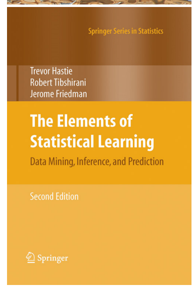
Course literature:

1. Christopher M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2006.
2. Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Second edition, Springer, 2009. (partly)

Recommended side reading:

1. Kevin P. Murphy. *Machine learning - a probabilistic perspective*, MIT Press, 2012.
2. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models Principles and Techniques*, MIT Press, 2012.
3. David Barber. *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.





A few words about probability distributions

- Important in their own right.
- Forms building blocks for more sophisticated probabilistic models.
- Touch upon some important statistical concepts.

See Chapter 2, Appendix B (useful summary) and Wikipedia.

The exponential family

The exponential family of distributions over x, parameterized by η,

$$p(x \mid \eta) = h(x)g(\eta)\exp\left(\eta^T u(x)\right)$$

Some of the members in the exponential family: Bernoulli, Beta, Binomial, Dirichlet, Gamma, Gaussian, Gaussian-Gamma, Gaussian-Wishart, Student's t, Multinomial, Wishart.

Multivariate Gaussian (I/VI)

$$\mathcal{N}(x; \mu, \Sigma) \triangleq \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Let us study a partitioned Gaussian,

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

with precision (information) matrix $\Lambda = \Sigma^{-1}$

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} = \begin{pmatrix} \Sigma_{aa}^{-1} + \Sigma_{aa}^{-1}\Sigma_{ab}\Delta_a^{-1}\Sigma_{ba}\Sigma_{aa}^{-1} & -\Sigma_{aa}^{-1}\Sigma_{ab}\Delta_a^{-1} \\ -\Delta_a^{-1}\Sigma_{ba}\Sigma_{aa}^{-1} & \Delta_a^{-1} \end{pmatrix}$$

where $\Delta_a = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$ is the Schur complement of Σ_{aa} in Σ.

Multivariate Gaussian (II/VI)

Theorem (Conditioning)

Let x be Gaussian distributed and partitioned $x = \begin{pmatrix} x_a & x_b \end{pmatrix}^T$, then the conditional density $p(x_a \mid x_b)$ is given by

$$\begin{aligned} p(x_a \mid x_b) &= \mathcal{N}(x_a; \mu_{a|b}, \Sigma_{a|b}), \\ \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}, \end{aligned}$$

which using the information (precision) matrix can be written,

$$\begin{aligned} \mu_{a|b} &= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b), \\ \Sigma_{a|b} &= \Lambda_{aa}^{-1}. \end{aligned}$$

Multivariate Gaussian (III/VI)

Theorem (Marginalization)

Let x be Gaussian distributed and partitioned $x = \begin{pmatrix} x_a & x_b \end{pmatrix}^T$, then the marginal density $p(x_a)$ is given by

$$p(x_a) = \mathcal{N}(x_a; \mu_a, \Sigma_{aa}).$$

Multivariate Gaussian (IV/VI)

Theorem (Affine transformations)

Assume that x_a, as well as x_b conditioned on x_a, are Gaussian distributed

$$p(x_a) = \mathcal{N}(x_a; \mu_a, \Sigma_a),$$
$$p(x_b \mid x_a) = \mathcal{N}(x_b; Mx_a + b, \Sigma_{b|a}),$$

where M is a matrix and b is a constant vector. The marginal density of x_b is then given by

$$p(x_b) = \mathcal{N}(x_b; \mu_b, \Sigma_b),$$
$$\mu_b = M\mu_a + b,$$
$$\Sigma_b = \Sigma_{b|a} + M\Sigma_a M^T.$$

Multivariate Gaussian (V/VI)

Theorem (Affine transformations, cont.)

The conditional density of x_a given x_b is

$$p(x_a \mid x_b) = \mathcal{N}(x_a; \mu_{a|b}, \Sigma_{a|b}),$$

with

$$\mu_{a|b} = \Sigma_{a|b}\left(M^T\Sigma_{b|a}^{-1}(x_b - b) + \Sigma_a^{-1}\mu_a\right)$$
$$= \mu_a + \Sigma_a M^T \Sigma_b^{-1}(x_b - b - M\mu_a),$$
$$\Sigma_{a|b} = \left(\Sigma_a^{-1} + M^T\Sigma_{b|a}^{-1}M\right)^{-1}$$
$$= \Sigma_a - \Sigma_a M^T \Sigma_b^{-1} M \Sigma_a.$$

Multivariate Gaussian (VI/VI)

Multivariate Gaussian's are important building blocks in more sophisticated models.

For more details, proofs and an example where the Kalman filter is derived using the above theorems is provided,

`www.control.isy.liu.se/student/graduate/MachineLearning/manipGauss.pdf`

Maximum Likelihood (ML) estimation

Maximum likelihood provides a systematic way of computing **point estimates** of the unknown parameters θ in a given model, by exploiting the information present in the measurements $\{x_n\}_{n=1}^N$.

Computing ML estimates of the parameters in a model amounts to:

1. Model the obtained measurements $x_1, \dots, x_N$ as a realisation from the stochastic variables $\mathbf{x}_1, \dots, \mathbf{x}_N$.
2. Decide on which model to use.
3. Assume that the stochastic variables $\mathbf{x}_1, \dots, \mathbf{x}_N$ are conditionally iid.

In ML the parameters θ are chosen in such a way that the measurements $\{x_n\}_{n=1}^N$ are as likely as possible, i.e.,

$$\widehat{\theta}^{\text{ML}} = \arg\max_{\theta} p(x_1, \cdots, x_N \mid \theta).$$

Bayesian modeling

The **goal** in Bayesian modeling is to compute the posterior $p(\theta \mid x_{1:N})$.

Provided that it makes sense from a modeling point of view it is convenient to choose prior distributions rendering a computationally tractable posterior distribution.

This leads to the so called **conjugate priors** (if the prior and the posterior have the same functional form, the prior is said to be a conjugate prior for the likelihood).

Again, only make use of conjugate priors if this makes sense from a modeling point of view!

Conjugate priors – example 1 (I/II)

Let $X = \{x_n\}_{n=1}^N$ be independent identically distributed (iid) observations of $x \sim \mathcal{N}(\mu, \sigma^2)$. Assume that the variance σ^2 is known.

The likelihood is given by

$$p(X \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right)$$

If we choose the prior as $p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$, the posterior will also be Gaussian. Hence, this Gaussian prior is a conjugate prior for the likelihood.

Conjugate priors – example 1 (II/II)

The resulting posterior is

$$p(\mu \mid X) = \mathcal{N}(\mu_B, \sigma_B^2),$$

where the parameters are given by

$$\mu_B = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}},$$

$$\frac{1}{\sigma_B^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

The ML estimate of the mean is

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

Conjugate priors – some examples

Likelihood	Model Parameters	Conjugate Prior
Normal (known mean)	Variance	Inverse-Gamma
Multivariate Normal (known mean)	Precision	Wishart
Multivariate Normal (known mean)	Covariance	Inverse-Wishart
Multivariate Normal	Mean and covariance	Normal-Inverse-Wishart
Multivariate Normal	Mean and precision	Normal-Wishart
Exponential	Rate	Gamma

Conjugate prior is just one of many possibilities!

Note that using a conjugate prior is **just one** of the many possible choices for modeling the prior! If it makes sense, use it, since it leads to simple calculations.

Let's have a look at an example where we do not make use of the conjugate prior and end up in a useful and interesting result.

Linear regression models the relationship between a continuous target variable t and an (input) variable x according to

$$\begin{aligned} t_n &= w_0 + w_1 x_{1,n} + w_2 x_{2,n} + \cdots + w_D x_{D,n} + \epsilon_n \\ &= w^T \phi(x_n) + \epsilon_n, \end{aligned}$$

where $\phi(x_n) = \begin{pmatrix} 1 & x_{1,n} & \dots & x_{D,n} \end{pmatrix}^T$ and $n = 1, \dots, N$.

Conjugate prior is just one of many possibilities!

Let $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$, resulting in the following likelihood

$$p(t_n \mid w) = \mathcal{N}(t_n \mid w^T \phi(x_n), \sigma^2).$$

Let us now assume w_n to be independent and Laplacian distributed (i.e. not conjugate prior), $w_n \sim \mathcal{L}(0, 2\sigma^2/\lambda)$

Def. (Laplacian distribution) $\mathcal{L}(x \mid a, b) = \frac{1}{2b} \exp\left(-\frac{|x-a|}{b}\right)$.

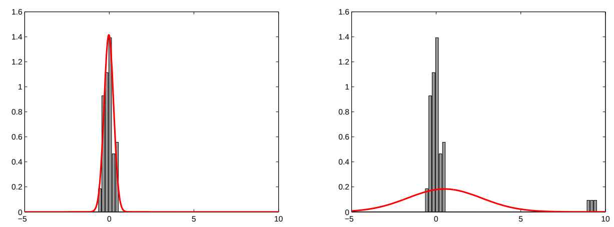
The resulting MAP estimate is given by,

$$w^{\text{MAP}} = \arg\max_{w} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2 + \lambda \sum_{n=1}^{D} |w_n|$$

Known as the **LASSO** and it leads to sparse estimates.

Robust statistics

Modeling the error as a Gaussian leads to very high sensitivity to outliers in the data. This is due to the fact that the Gaussian assigns very low probability to points far from the mean. The Gaussian is said to have *"thin tails"*.



Two possible solutions

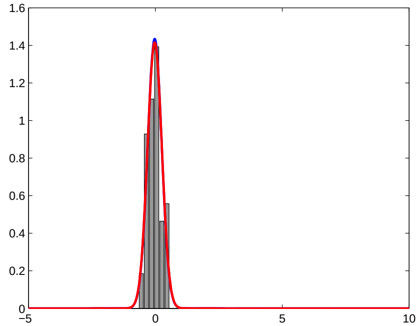
1. Model using a distribution with "heavy tails".
2. Outlier detection models

Example: heavy tails (I/III)

Generate $N = 50$ samples,

$$x \sim \mathcal{N}(0, 0.1)$$

Plot showing a realization (gray histogram) and the corresponding ML estimate of a Gaussian (red) and a Student's t-distribution (blue).



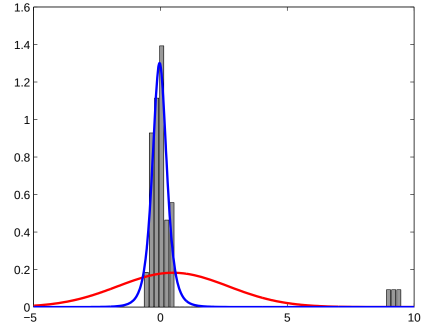
Note that (as expected?) the red curve sits on top of the blue curve.

Example: heavy tails (II/III)

Let us now add 3 outliers $9, 9.2$ and 9.5 to the data set.

Plot showing resulting ML estimate of a Gaussian (red) and a Student's t-distribution (blue).

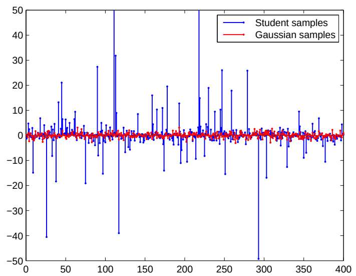
Clearly the Student's t-distribution is a better model here!

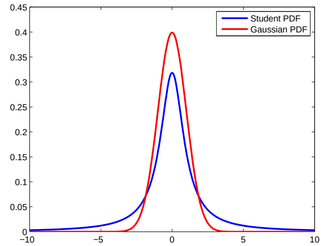


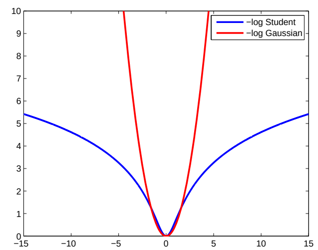
Example: heavy tails (III/III)

Below: 400 samples from a Student's t-distribution and a Gaussian distribution.

Right: The corresponding pdf's and negative log-likelihoods.







Outlier detection models

Model the data as if it comes from a mixture of two Gaussians,

$$\begin{aligned} p(x_i) &= p(x_i \mid k_i = 0)p(k_i = 0) + p(x_i \mid k_i = 1)p(k_i = 1) \\ &= \mathcal{N}(0, \sigma^2)p(k_i = 0) + \mathcal{N}(0, \alpha\sigma^2)p(k_i = 1). \end{aligned}$$

where $\alpha > 1$, $p(k_i = 0)$ is the probability that the sample is OK and $p(k_i = 1)$ is the probability that the sample is an outlier.

Note the similarity between these two "robustifications":

- The Student's t-distribution is an infinite mixture of Gaussians, where the mixing is controlled by the ν-parameter.
- The outlier detection model above consists of a sum of two Gaussians.

Summary – robust statistics

- Do not use distributions with thin tails (non-robust) if there are outliers present. Use more realistic robust "heavy tailed" distribution such as the Student's t-distribution or simply a mixture of two Gaussians.
- A nice account on robustness in a computer vision context is available in Section 3.1 in

  B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. **Bundle Adjustment - A Modern Synthesis**. In: *Vision algorithms: theory and practice.* Lecture Notes in Computer Science, Vol 1883:152–177. Springer, Berlin, 2000. dx.doi.org/10.1007/3-540-44480-7_21

Example - range measurements with outliers

We measure range (r), contaminated by a disturbance $d_n \geq 0$ and noise $e_n \sim \mathcal{N}(0, \sigma^2)$, $y_n = r + d_n + e_n$. Compute the MAP estimate of $\theta = \{r, d_1, \dots, d_N\}$ under an exponential prior on d_n,

$$p(d_n) = \begin{cases} \lambda \exp(-\lambda d_n), & d_n \geq 0, \\ 0, & d_n < 0. \end{cases}$$

Resulting problem

$$\widehat{\theta}^{\text{MAP}} = \arg\max_{\theta} p(\theta \mid y_{1:N}) = \arg\min_{\theta} \sum_{n=1} N \frac{(y_n - r - d_n)^2}{\sigma^2} + \lambda \sum_{n=1}^{N} d_n$$

For details, see Example 2.2. in the PhD thesis of Jeroen Hol.

This principle is used for ultra-wideband positioning, incorporated into MotionGrid (www.xsens.com/en/general/motiongrid) from our partners Xsens (www.xsens.com).

Important message!

Given the computational tools that we have today it can be rewarding to resist the Gaussian convenience!!

We will try to repeat and illustrate this message throughout the course using theory and examples.

A few concepts to summarize lecture 1

Supervised learning: The data consists of both input and output signals (e.g., regressions and classification).

Unsupervised learning: The data consists of output signals only (e.g., clustering).

Reinforcement learning: Finding suitable actions (control signals) in a given situation in order to maximize a reward. (Very similar to control theory)

Conjugate prior: If the posterior distribution is in the same family as the prior distribution, the prior and posterior are *conjugate distributions* and the prior is called a conjugate prior for the likelihood.

Maximum likelihood: Choose the parameters such that the observations are as likely as possible.