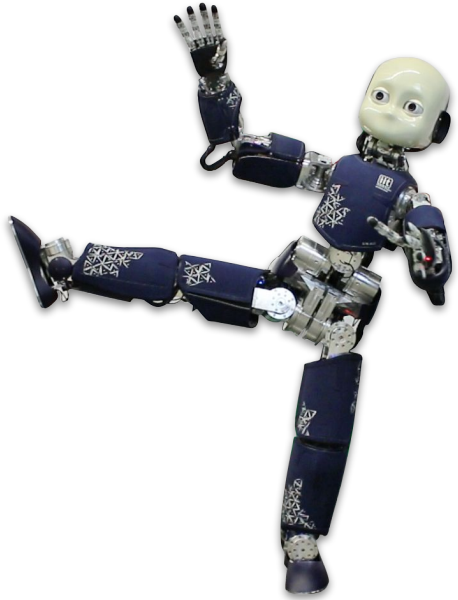
Bayesian Feature Regression (L9)

Machine Learning (190.012)

Univ.-Prof. Dr. Elmar Rueckert

Telefon: +43 3842 402 - 1901
Email: teaching@ai-lab.science



WO AUS FORSCHUNG ZUKUNFT WIRD

Chair of Cyber-Physical-Systems

Bonus Point Quiz

Get **all** answers right to receive 2 Bonus Point!

Link to session

https://tweedback.de/u4v6

Which statements about the KL-divergence are correct.

Multiple answers are possible

Vote options:

- The KL-divergence is a deterministic distance measure.
- The KL-divergence is an asymmetric distance measure.
- The KL-divergence utilizes the log function to put equal weights on the fractals, e.g., f(Q(A) / P(A)) = -f(P(A) / Q(A))
- The mode seeking KL divergence KL(q||p) focuses on a single mode in a multi-modal data distribution p.
- The moment matching KL divergence KL(q||p) focuses on a single mode in a multi-modal data distribution p.

Which statements about basis functions in feature regression are correct.

Multiple answers are possible

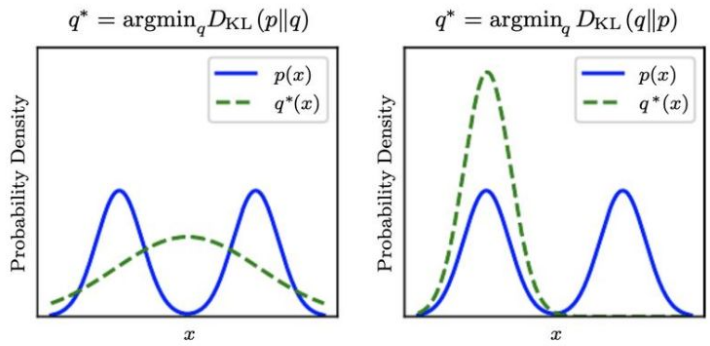
Vote options:

- We introduced basis functions to transform the outputs 'y' in linear regression.
- Polynomial basis functions can be defined as y = w1 + w2 x + w3 x^2 + w4 x^3 ...
- Basis functions in linear regression transform the inputs into a feature space.
- Basis functions were introduced to also implement models with more parameters than the input dimensions.
- Basis functions were introduced to add noise to the model predictions.

Recap the KL-Divergence concepts

Univ.-Prof. Dr. Elmar Rueckert
08.05.2024

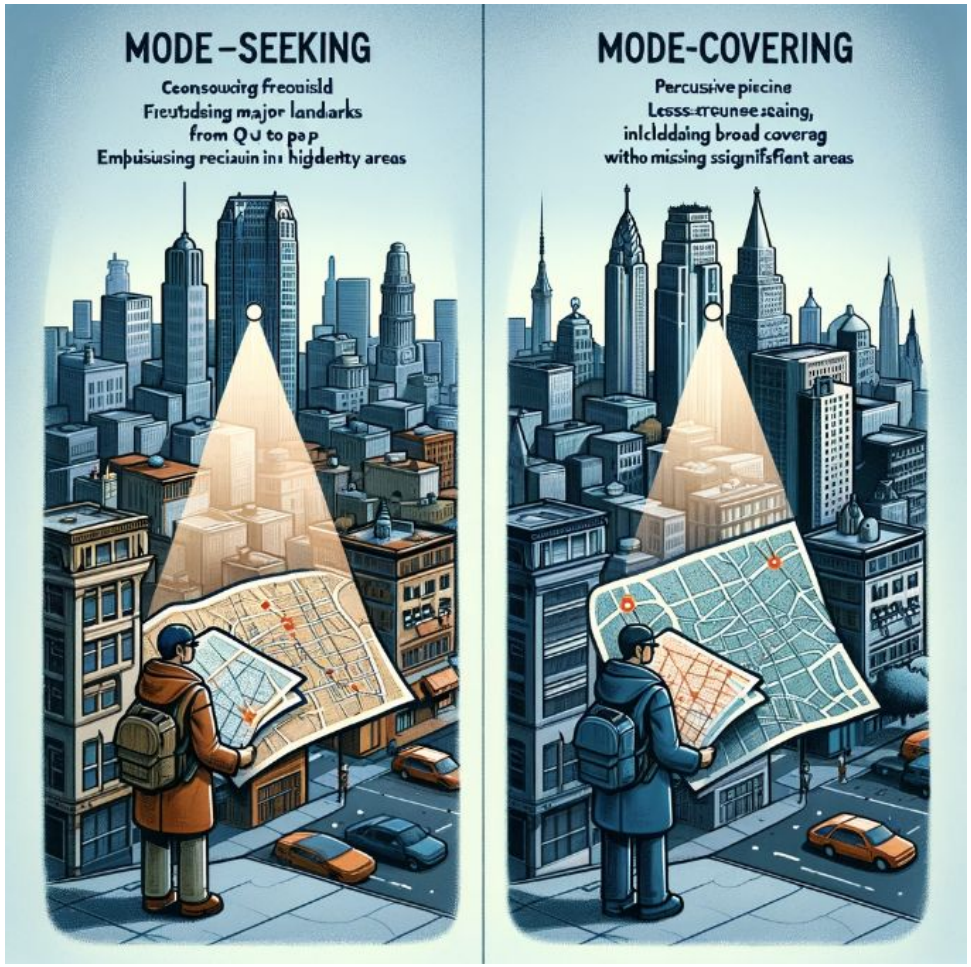
Chat-GPT4 Prompt on teaching the KL divergence concepts



User
I am teaching a course on 'introduction to machine learning'. Last week I explained the KL-divergence and motivated the usage of the log function to equalize the fractals of individual events, i.e. log (P(A)/Q(A)) = - log(Q(A)/P(A)). Now what makes the difference between the mode seeking and the moment matching KL divergence is the importance weighting using either P(A) log(P(A)/Q(A)) or Q(A) log(Q(A)/P(A)).
I visually illustrated that drawing the distributions P(X) and Q(X). Can you provide an alternative explanation, maybe in an understandable intuitive way?

ChatGPT

Absolutely, let's delve into an intuitive explanation of the concepts you've covered—KL divergence, particularly the distinctions between its two forms, and how they relate to mode seeking and moment matching.
Understanding KL Divergence



https://cloud.cps.unileoben.ac.at/index.php/s/dMGqWc2RSf68pQt

Outlook of this lecture

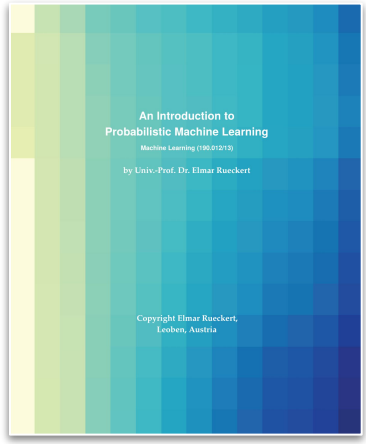
Probability Theory

PROBABILISTIC REGRESSION 12

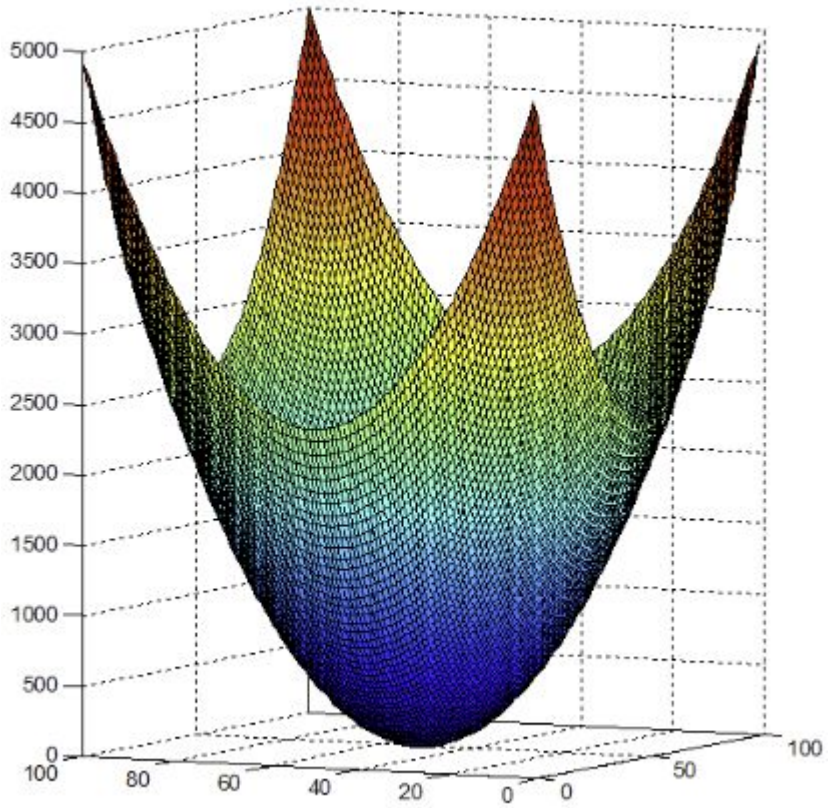
3 Linear Probabilistic Regression 13
3.1 Linear Feature Regression 13
3.2 Basis function models 14
3.3 Logisitic Regression 15
3.4 Maximum Likelihood 15

Note that the lecture will take place in the lecture room. However, I will not use slides. Instead I will **derive least squares feature regression and maximum likelihood solutions** on the black board.

Have a look at the pages 13-16 in the ML book (https://cloud.cps.unileoben.ac.at/index.php/s/iDztK2ByLCLxWZA).



Motivation for Feature Transformations



Linear Models and Quadratic Objectives are 'nice'!

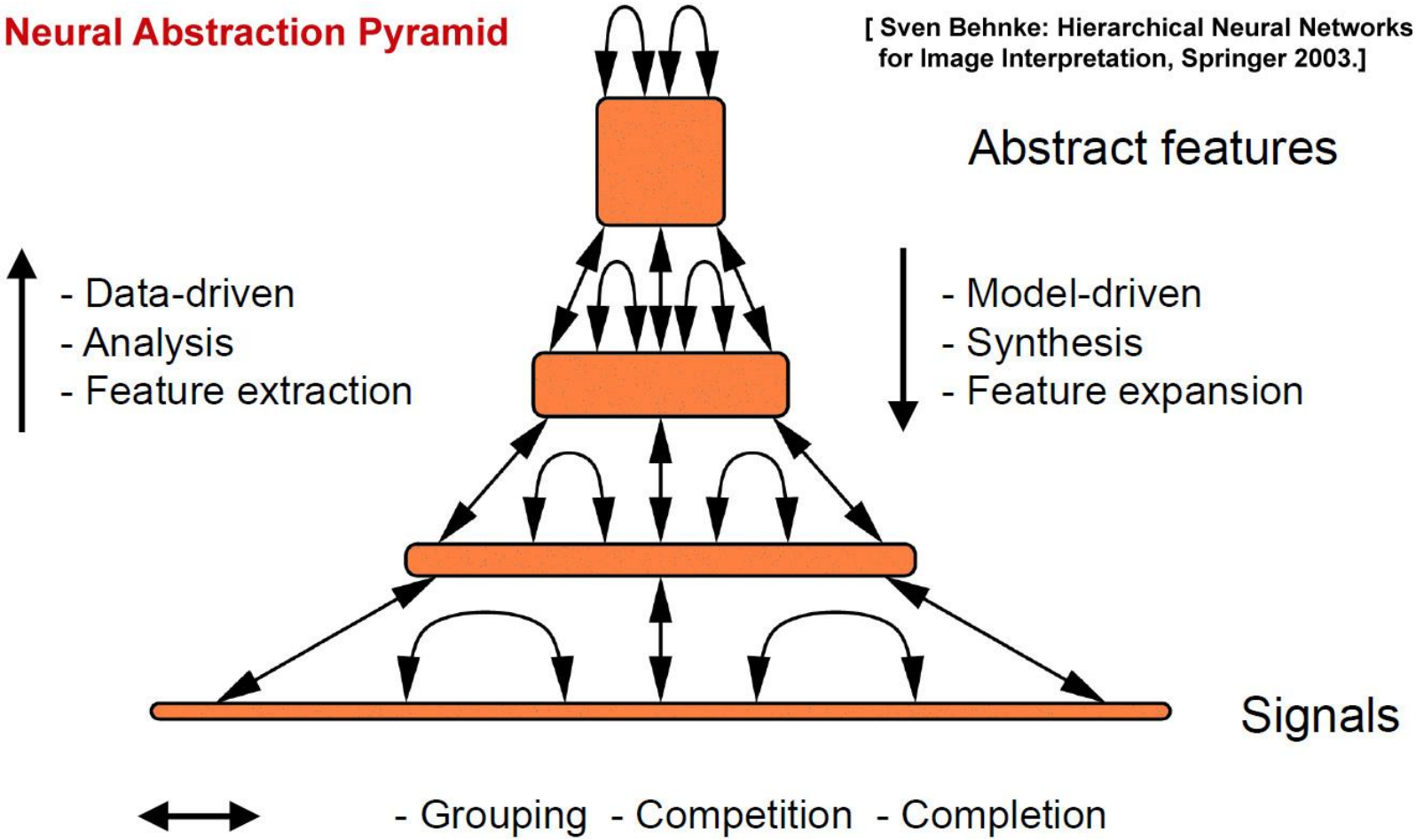
- **They have a single global minima!**
- Thus, the optimal solution can be computed in a single update step, e.g., through least squares regression!

- Note that most objectives are non-convex and most models are non-linear. They require iterative update mechanisms like iterative gradient descent!

One major goal of a feature transformation is to use linear models with convex objectives **in a feature space**. The feature space mapping introduces a non-linearity that increases the model complexity and thus its data generation/reproduction capabilities.

Motivation for Feature Transformations

Another major goal of feature transformations are abstractions to ‘explain’ the data!



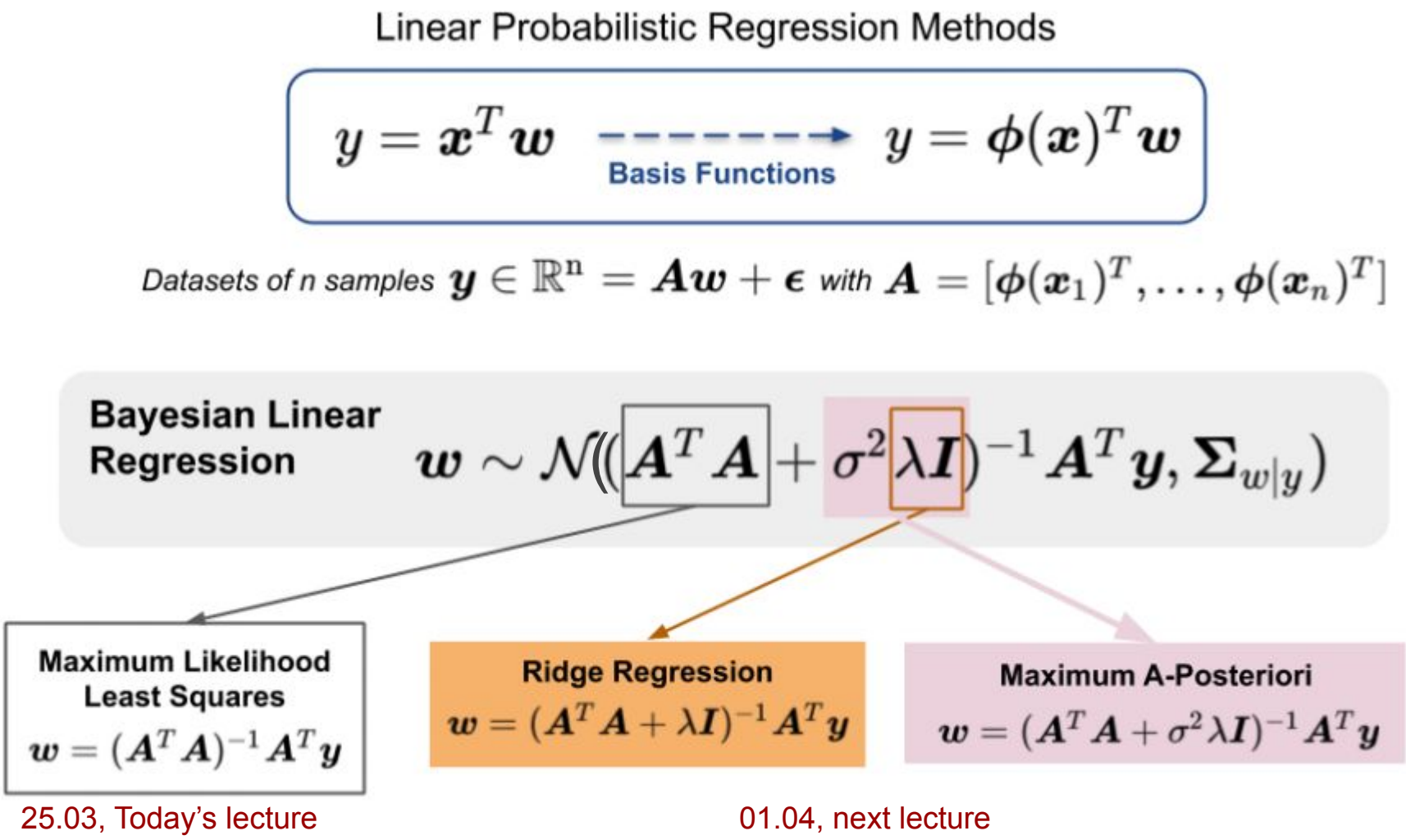
Regression Models Discussed in this Course

Linear Probabilistic Regression Methods

$$y = \boldsymbol{x}^T \boldsymbol{w} \xrightarrow[\text{Basis Functions}]{} y = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{w}$$

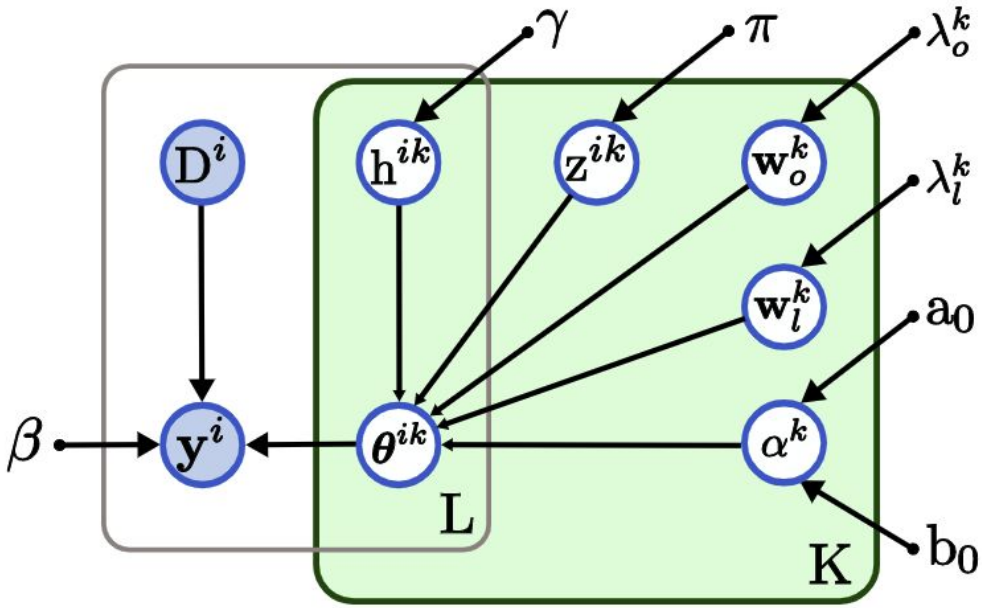
Datasets of n samples $\boldsymbol{y} \in \mathbb{R}^{\mathrm{n}} = \boldsymbol{A}\boldsymbol{w} + \boldsymbol{\epsilon}$ *with* $\boldsymbol{A} = [\boldsymbol{\phi}(\boldsymbol{x}_1)^T, \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n)^T]$

Regression Models Discussed in this Course



Research Studies exploiting such Bayesian Models

Rueckert, Elmar; Mundo, Jan; Paraschos, Alexandros; Peters, Jan; Neumann, Gerhard
Extracting Low-Dimensional Control Variables for Movement Primitives Inproceedings
In: Proceedings of the International Conference on Robotics and Automation (ICRA), 2015.
Links | BibTeX | Tags: movement primitives, Probabilistic Inference



Thank you for your attention!

Univ.-Prof. Dr. Elmar Rückert

Chair of Cyber-Physical-Systems

Montanuniversität Leoben

Franz-Josef-Straße 18,

8700 Leoben, Austria

Phone: +43 3842 402 – **1901** (Sekretariat CPS)

Email: teaching@ai-lab.science

Web: https://cps.unileoben.ac.at