University of Luebeck
Institute for Robotics and Cognitive Systems
Prof. Dr. Elmar Rueckert

Probabilistic Machine Learning (RO5601)

February 17th, 2020

Name : ....
Matriculation number : ....
Signature : ....

Question	Points	Score
1	25	
2	25	
3	25	
4	25	
Total:	100	

1 General Comments (read them carefully)

- Check at first if you got all pages of the exam. The exam includes a cover page and printouts with the page numbers ranging to 15.
- Start by filling out the cover page.
- Your answers and results need to be placed right after the questions part. If you need more space, leave a solid dot or star symbol and ask for an additional sheet.
- List all assumptions if you make any to get to your result.
- Your answers should be short and concise. Keyword style is also allowed.
- All aids with the exception of electronic devices are permitted.
- Non-programmable calculators are allowed.
- Question are marked indicating the complexity, i.e., easy (•), medium (• •) and hard (• • •).
- The processing time for the exam is 90 minutes.

Good luck!

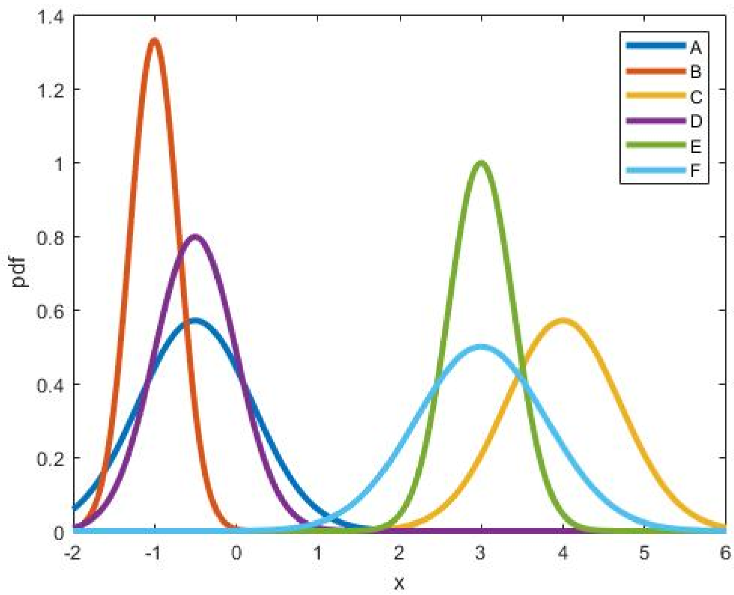


Figure 1: Gaussian Distributions

Introduction to Probability Theory

1. (25 Points)

   (a) • (3 Points)

   Please allocate the probability density functions to the correct set of parameters by writing the correct letter into the boxes below.

   Gaussian Distribution:

$$\mathcal{N}(x|\mu,\sigma) := \frac{1}{(2\pi\sigma^2)^{1/2}} \exp^{\{-\frac{1}{2\sigma^2}(x-\mu)^2\}}.$$

μ	3	3	4	-1	-0.5	-0.5
σ	0.8	0.4	0.7	0.3	0.7	0.5
Function Letter						

(b) • (4 Points)

Below is a Venn Diagram depicted which shows the sport preferences (A: Soccer, B: Baseball) of 50 children. Please determine $p(A \cap B)$, $p(A \cup B)$, $p(A)$ and $p(B)$.

Hint: A child can like both, soccer and baseball.

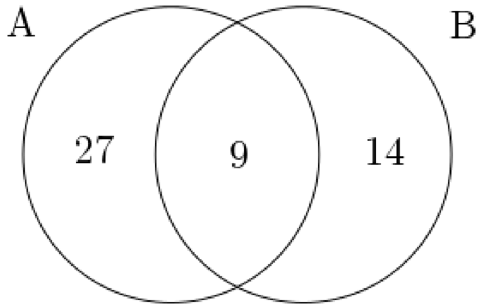


Figure 2: Gaussian Distributions

(c) • • (9 Points)

Imagine you are going to a blood donation. For security reasons a blood probe for every participant is checked for Hepatitis B. The control of your probe using a standard Hepatitis B test gives back a positive result.

A standard Hepatitis B test is 99.7% sensitive and 99.5% specific. Thus it gives back a positive result with a probability of 99.7% for infected persons and a negative result with a probability of 99.5% for non-infected persons. About 0.2% of the German population is infected by Hepatitis B. State the Bayes' Theorem and name all parts (Likelihood, posterior, prior) properly! Give an example why the Bayes' Theorem is useful for robotic applications! Calculate the probability that you are really infected with Hepatitis B using the Bayes' Theorem!

(d) • • • (9 Points)

You are in a casino and observe a player who always bets on the number 13 by Roulette. In 5 games he won 3 times. You suspect that the player cheats. Please use the Bayesian View to either verify or disprove your suspicion. We thereby assume the prior probability density distribution to be uniformly distributed over the interval $[0, 1]$. As trust interval we use $[0, \frac{2}{37}]$.

Hint: The probability of winning by betting on the number 13 is $p_{13} = \frac{1}{37}$.

Linear Probabilistic Regression

2. (25 Points)

   (a) • • (5 Points)

   Given are the scalar a, the vectors $\boldsymbol{g} \in \mathbb{R}^n$, $\boldsymbol{\mu} \in \mathbb{R}^M$ and the matrix $K \in \mathbb{R}^{n \times M}$. Compute the derivative of the term

$$\frac{\delta}{\delta\mu}(\boldsymbol{\mu}^T \boldsymbol{K}^T \boldsymbol{g} + a)$$

   (b) • • (5 Points)
   Given is a linear system of the form

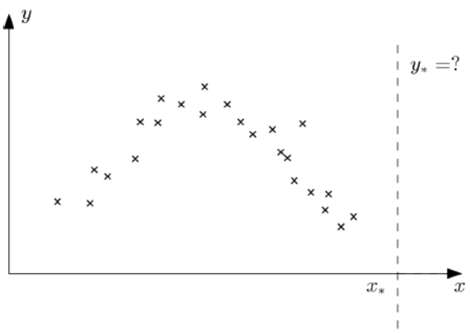
$$\boldsymbol{K}\boldsymbol{\mu} \;=\; \boldsymbol{g}.$$

   Derive the solution for $\boldsymbol{\mu} = ...$ List also intermediate steps.

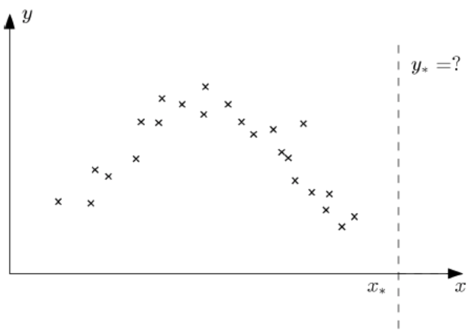
(c) • (4 Points)
Given is again the linear system of the form $\boldsymbol{K}\boldsymbol{\mu} = \boldsymbol{g}$.
Write down the ridge regression result for $\boldsymbol{\mu} = ...$

Explain in 1-2 sentences why ridge regression is useful here.

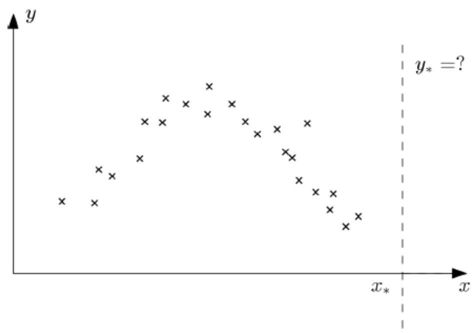
(d) • (3 Points)
Illustrate overfitting, underfitting and the optimal fit based on the examples in Figure 3. Mark the prediction y_* by adding a circle or cross to the three panels.



(a) **Overfitting**



(b) **Underfitting**



(c) **Optimal Fit**

Figure 3: Illustration of a one dimensional temperature data set. Here, $y = f(x)$.

(e) • • • (8 Points)

Compute the derivative of $\frac{\delta}{\delta \boldsymbol{a}}\{\log \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{a}, \lambda \boldsymbol{I}\right)\}$. List intermediate steps.

Nonlinear Probabilistic Regression

3. (25 Points)

   (a) • (6 Points)

   Below you find a sketch of a Gaussian Process estimating the amount of oil ($y = y(x)$) based on the measured values marked by a dot. The aim was to estimate the amount of oil for all positions x by drawing the mean (dotted line) and the variance (solid lines). Also, the next drilling point for searching for a maximum should be drawn. However, some mistakes have crept in. Please mark the mistakes in the sketch (do not use red) and shortly explain each mistake in the box below.

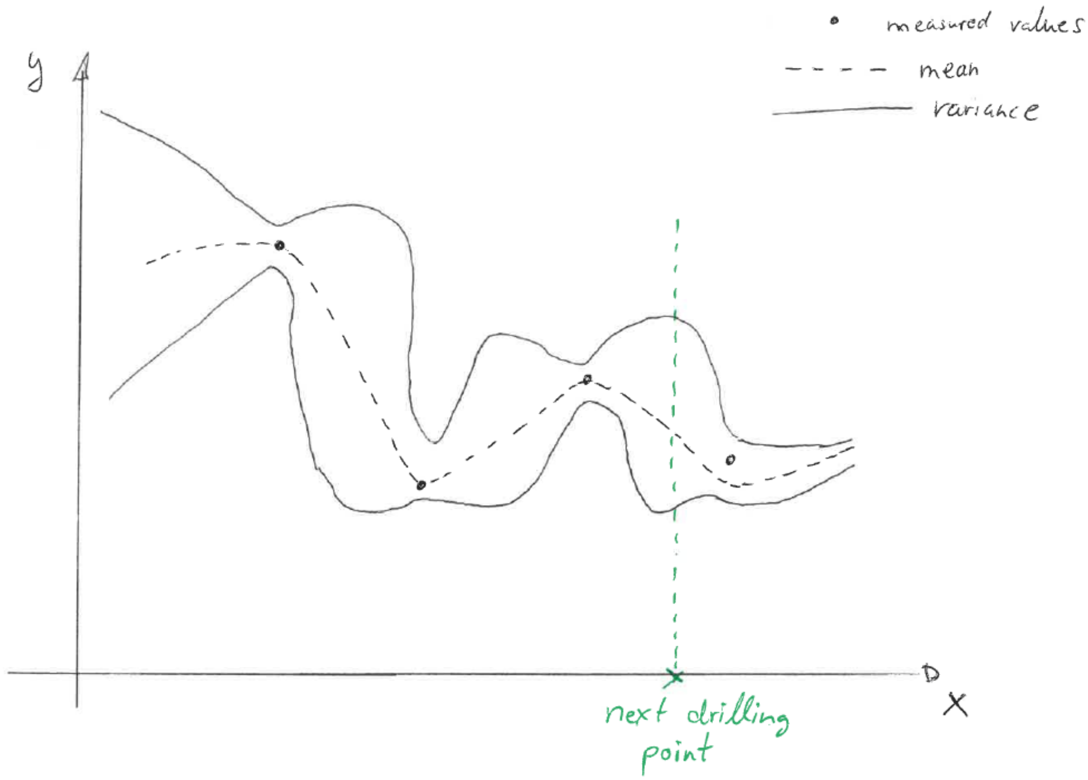


Figure 4: Oil Drilling GP Sketch

(b) • • (10 Points)

Mark the correct statements about GPs and GPs with built-in GMRFs. Note you can get +1 point for a correct answer and −1 point for a wrong choice. The minimum is zero points.

- ◯ The computational time of GPs is always less than the computational time of GPs with built-in GMRFs
- ◯ GPs require a definition of a regular grid, the so called generating points
- ◯ The covariance matrix $\boldsymbol{K}$ used for GPs has to be positive definite
- ◯ GPs are linear regression models which can be used for modeling the dynamics of robotics systems
- ◯ Let $\boldsymbol{x}_i, y_i$ for $i = 1, \ldots, n$ the training data for the GPs, then $\boldsymbol{K}_{ij} = \text{kernel}(x_i, x_j)$ defines a required covariance matrix if kernel is a valid kernel function
- ◯ The hyper-parameters for GPs with built-in GMRFs can be learned using the Log-Likelihood method
- ◯ The covariance matrix $\boldsymbol{K}$ for GPs has to be strictly diagonal dominant with $\boldsymbol{K}_{ii} > 0$, $\forall i$.
- ◯ GPs with built-in GMRFs have the computational complexity of $\mathcal{O}(NM^2)$, where M is the number of generating points and N the number of measurements
- ◯ Gaussian processes are stochastic processes
- ◯ GPs have the computational complexity of $\mathcal{O}(N^4)$, where N is the number of measurements

(c) • • (4 Points)

We computed for $x \in [0, 5]$ the covariance function of a Gaussian Process using the four different Kernels:

$$k1(x1, x2) = \exp\left(-\frac{||x_1 - x_2||^2}{2}\right) ,$$
$$k2(x1, x2) = 10 \cdot \exp\left(-\frac{||x_1 - x_2||^2}{20}\right) ,$$
$$k3(x1, x2) = \left(x_1^\top x_2 + 2\right) ,$$
$$k4(x1, x2) = \exp\left(-|x_1 - x_2|\right) ,$$

We then sampled a set of 10 different latent functions f according to the Gaussian process prior:

$$f \sim \mathcal{N}(0, K) ,$$

for all four different Kernel/Covariance Matrices and plotted them. Please allocate the kernel functions to the correct plots by writing $k1, \ldots, k4$ above the plots.

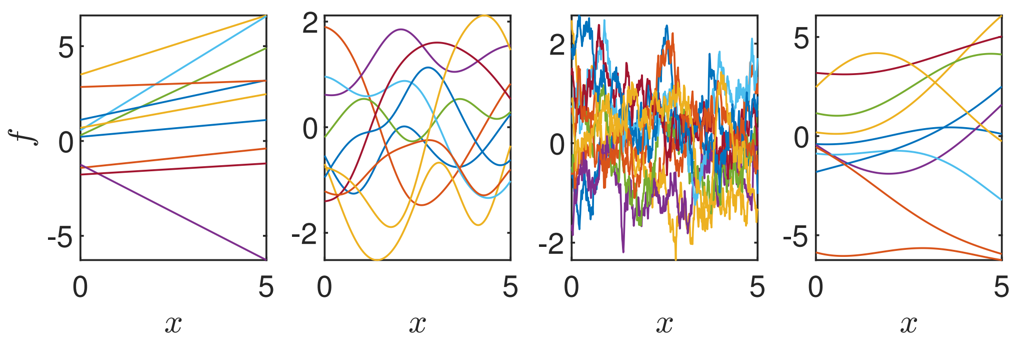


Figure 5: Kernel Function Plots

(d) • • (5 Points)

For GPs with built-in GMRFs we proofed that the precision $Q_{ij} = 0$ if i, j are not neighbored vertices. Write down the complete precision matrix for a CAR(1) model on a regular 3×2 lattice. Therefore, first sketch the regular lattice and number the vertices.

Hint: The q-patterns for a CAR(1) model are given below:

General q-pattern:

$$q = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 + \kappa^2 & -1 \\ 0 & -1 & 0 \end{bmatrix} .$$

Boundary q-pattern:

$$q_{\text{edge}} = \begin{bmatrix} -1 & 0 \\ 3+\kappa^2 & -1 \\ -1 & 0 \end{bmatrix},$$

Corner q-pattern:

$$q_{\text{corner}} = \begin{bmatrix} 2+\kappa^2 & -1 \\ -1 & 0 \end{bmatrix}.$$

Probabilistic Inference

4. (25 Points)

   (a) • • (8 Points)
   Mark the correct statements about *probabilistic time series models (PTSM).* Note you can get +1 point for a correct answer and −1 point for a wrong choice. The minimum is zero points.

   - ◯ A coupled PTSM is a model where all parameters are known
   - ◯ A coupled PTSM encodes the correlation between individual dimensions or joints
   - ◯ Radial basis functions *cannot* be used to approximate any non-linear function
   - ◯ Radial basis functions in PTSMs model the movement phase
   - ◯ The bandwidth h parameter in radial basis functions defines the support or width of the Gaussians
   - ◯ The integral over the product of the two dependent Gaussian distributions, i.e., $\int \mathcal{N}(x|\boldsymbol{S}\boldsymbol{y}, \sigma^2)\ \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\ d\boldsymbol{y}$, *cannot* be solved analytically
   - ◯ If the number of basis functions in PTSMs is increased, the model complexity increases
   - ◯ PTSMs can be used to predict unobserved dimensions from exploiting the correlation of multi-dimensional systems

   (b) • • (6 Points)
   Given is a probabilistic trajectory model represented by the parameter vector $\boldsymbol{\theta}$. The distribution over trajectories $\boldsymbol{\tau}$ is defined as

$$p(\boldsymbol{\tau}) \quad = \quad \int p(\boldsymbol{\tau}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}. \tag{9}$$

   Show that the model in Equation 9 can be derived by using the definition of the *joint distribution* over $p(\boldsymbol{\tau}, \boldsymbol{\theta})$ and by applying the *Marginal rule.* Clearly mark when the these rules are used in your derivations.

(c) ● ● ● (8 Points)

Now use the Gaussian distributions for $p(\boldsymbol{\tau}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\tau}|\boldsymbol{A}\boldsymbol{w}, \sigma_y^2\boldsymbol{I})$ and $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{w|y}, \boldsymbol{\Sigma}_{w|y})$ and derive the analytic solution for $p(\boldsymbol{\tau})$.

Solve the integral using the Gaussian identity

$$\int \mathcal{N}(\boldsymbol{x}|\boldsymbol{a} + \boldsymbol{F}\boldsymbol{y}, \boldsymbol{A})\mathcal{N}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{B})d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x}|\boldsymbol{a} + \boldsymbol{F}\boldsymbol{b}, \boldsymbol{A} + \boldsymbol{F}\boldsymbol{B}\boldsymbol{F}^T).$$

Derive the solution for $p(\boldsymbol{\tau})$ and list intermediate steps.

(d) • (2 Points)

Given are the trajectories of right wrist motions of human subjects. In Figure 6 the y-coordinate of the trajectories are shown.

Sketch how the learned trajectory distribution in Equation 9 would model the data. (1) First, illustrate the mean estimate by a solid line. (2) Second, mark the model standard deviation by drawing *error bars* or by drawing a *shaded region.*

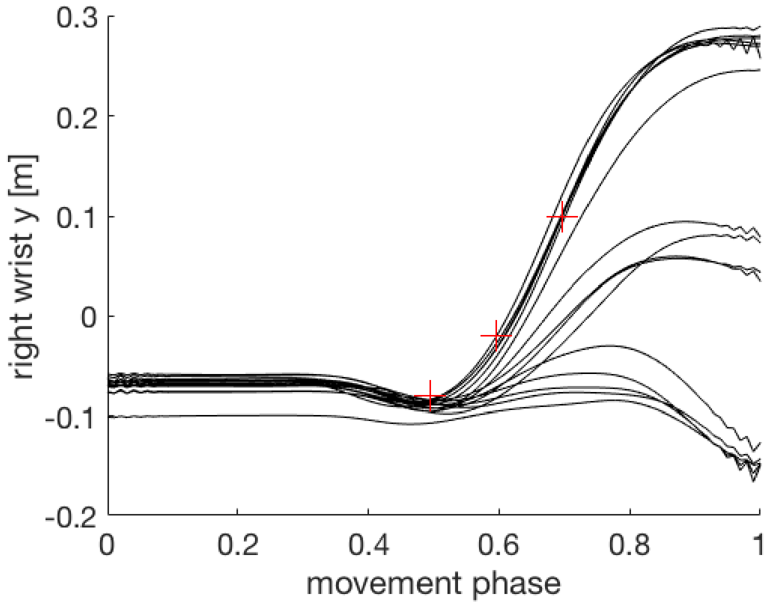


Figure 6: Illustration of y-coordinates of right wrist motions of human subjects.

(e) • (1 Point)

Sketch a trajectory prediction with *high* observation noise ($\sigma_o = 1$). Given are the three observations marked by red crosses in Figure 6. Hint: adding a line denoting the prediction is sufficient.