1 Lecture 1

Why Set ∞ to 0?

Prove that the solution for $\boldsymbol{x}$ on the corresponding slide has the smallest norm of all solutions. Also prove that for all $\boldsymbol{b}$ which are not in the range of $\boldsymbol{A}$, the above solution is optimal in a least squares sense.

Sherman-Morrison-Woodbury

Prove that the Sherman-Morrison-Woodbury formula is correct.

Matrix Functions

Find a compact expression for vec $\boldsymbol{x}\boldsymbol{x}^T$.

Avoiding Integration

Calculate the expectation of x under a truncated normal distribution:

$$\int_{x=\alpha}^{\beta} x(2\pi)^{-0.5}\lambda^{0.5}\frac{1}{\Phi(\beta;\mu,\lambda)-\Phi(\alpha;\mu,\lambda)}\exp(-0.5\lambda(x-\mu)^2)dx$$

What does Φ represent here? Hint: Think about the "normalisation property" of all density functions.

Citations

Further information about linear algebra can be found in [7], [6] and [4]. Information about matrix differential calculus (including many pre-calculated relations) is available in [5]. A good reference for mathematical functions is in [1].

2 Lecture 2

Likelihood and Linear Regression

Derive the maximum likelihood estimates for parameters $\boldsymbol{\theta}$ and λ. Find a *qualitative interpretation* of the size of λ in dependence of the prediction error.

Classification in the Sampling Paradigm

Assume a two class problem. Class 1 has prior probability π and a class conditional density which is a multivariate d-dimensional Gaussian with covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}_1$. Class 0 has a class conditional density which is a multivariate d-dimensional Gaussian with covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}_0$. Answer the following questions:

- What is the class prior of class 0?
- Show that the posterior probability is given by a logistic distribution.
- Express the regression parameter $\boldsymbol{\theta}$ in terms of the parameters of the class conditional densities and class priors.

Data Transformations

Assume that you wish to apply some data analysis tool to a data-set, where the independent variables are given as matrix $\boldsymbol{X}$, such that the n-th sample is the n-th row vector $\boldsymbol{X}[n, .]$. Many data analysis tools might benefit from data whitening. This term is used for two different kinds of transformations. Both approaches remove the overall location from the data (i.e. the data is transformed such that the sample mean is zero). The methods differ in the way they treat the scale of the data.

- One approach treats each variable (column in $\boldsymbol{X}$) independently and adjusts the sample variance such that the transformed data has unit standard deviation.
- The other approach rotates the data such that all covariances of the transformed data are zero and "stretches" the new coordinate system such that the largest variance component has unit standard deviation.

Develop sensible mathematical expressions which describe both transformations and implement a MatLab function "data_whiten" which expects two parameters, one contains the data, the other a mode flag which defines which transformation will be applied. Discuss advantages and disadvantages of both methods.

N-fold Cross Testing

Implement a MatLab function for foldspit.

Citations

An overview about data analysis which goes far beyond the topics discussed in lecture 2 is provided by [3].

3 Lecture 3

Collection of Density Functions

The following definitions of probability distributions may be used during the exam.

Univariate and multivariate Gaussian distribution:

$$p(x|\mu, \lambda) = (2\pi)^{-\frac{1}{2}}|\lambda|^{\frac{1}{2}} \exp(-0.5\lambda(x-\mu)^2)$$

$$p(\boldsymbol{x}|\boldsymbol{\mu}\boldsymbol{\Lambda}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp(-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}))$$

Gamma distribution:

$$p(\lambda|g, h) = \frac{h^g}{\Gamma(g)}|\lambda|^{(g-1)} \exp(-h\lambda)$$

Multinomial one distribution:

$$P(I|\pi) = \prod_{k=1}^{K} \pi_k^{\delta(I=k)}$$

Dirichlet Distribution:

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Note! for $K = 2$ we get the Bernoulli distribution as special case of the Multinomial-one distribution:

$$P(I|\pi) = \pi^{\delta(I=1)}(1-\pi)^{\delta(I=2)}$$

and the Beta distribution as a special case of the Dirichlet distribution:

$$p(\pi) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}$$

Student-t distribution:

$$p(\mu|\theta, \kappa, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}|\kappa|^{0.5}(\nu\pi)^{-0.5}\left(1 + \frac{(\mu-\theta)^2\kappa}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Bayesian Linear Regression

- Derive all equations you need for the marginal posterior distribution over regression coefficients, the marginal likelihood and the predictive distribution.

- Implement a MatLab function which calculates the marginal likelihood of a regression model and a function which provides the predictive distribution of a novel test sample.

We will need the MatLab implementation for a data analysis task in the practical session. It is strongly recommended that you start this implementation soon, such that you have the code ready in December!

Citations

An introduction to Bayesian data analysis can be found in [8] a thorough discussion of Bayesian Theory is provided in [2].

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover, New York, 1965.

[2] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, Second Edition.* Wiley, New York, 2000.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, second edition, 1989.

[5] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus.* Wiley, New York, 1988. paper back in 1999.

[6] C. D. Meyer. *Matrix analysis and applied linear algebra.* SIAM, 2000.

[7] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C++.* Cambridge University Press, Cambridge, second edition, 2002.

[8] C. P. Robert. *The Bayesian choice, a decision theoretic motivation.* Springer, New York, 1994.