The Neural Code

Summer term 2022, Christian Leibold, Yuk-Hoi Yiu

Problem Set 1 **—Maximum Likelihood, Regularization—** **21.06.2021**

Due date: 30.06.2021

---

1. (Pseudoinverse) Consider the linear equation $y = A\,x$ in which A is a non-invertible matrix.

(a) Show that the pseudo inverse matrix

$$A^* = (A^T A)^{-1} A^T$$

minimizes the risk function

$$R = \sum_{i=1}^{n} |AA^*\,y_i - y_i|^2 \ .$$

for a set of observations $y_1, \ldots, y_N$.

(b) Interpret the loss function in terms of the estimates $\hat{x}_i = A^* y_i$.

(c) Compute the pseudoinverse of the 4×6 matrix
`A=np.array([[1,2,-2,4,0,-1],[0,0,2,-1,2,1],[9,0,-1,0,1,1],[2,1,-2,0,2,0]])`.
Analyse the spectrum (eigenvalues) of the matrix $(A^T A)^{-1}$. What's the problem?

(d) Generate a 6×100 normally distributed random matrix x (100 input vectors) and derive the observables $y = A\,x$. Compute the risk for $A^* = (A^T A)^{-1} A^T$ and the approximate unity $A\,A^*$. What's the problem?

(e) Introduce the regularization parameter $\delta > 0$ to define

$$A^*_\delta = (A^T A + \delta\, E)^{-1}\, A^T \ ,$$

with E denoting the unity matrix and compute the risk and $A\,A^*$ as a function of δ. Interpret the result!

2. (General linear model)

Measurements $y^{(i)}, i = 1, \ldots, N$ are assumed to linearly depend on M regressors $x_1^{(i)}, \ldots, x_M^{(i)}$, i.e.,

$$y^{(i)} = \mathbf{w} \cdot \mathbf{x}^{(i)}$$

(a) Load the two data sets `olsx.txt` and `olsy.txt` as example data points and regressors and fit a general linear model using `statsmodels.regression.linear_model.OLS`.

(b) Interpret the output of the `summary` method in `OLS.fit()`.

(c) As a next step, explore whether there are interactions between input dimensions, by including all second order monomials $x_n\,x_m$ as regressors.

3. (Density Estimation)

Consider the stochastic variable G for which we do not know the underlying probability distribution p_G. We guess that p_G can be approximated by

$$q_G(g) = \lambda \, \frac{(g\,\lambda)^a}{a!} \, e^{-g\lambda} \, \mathcal{H}(g)$$

with $\lambda > 0$ and $a = 0, 1, 2, ...$ being a non-negative integer. $\mathcal{H}$ is the Heaviside step function.

(a) Assume you have collected K data samples $\gamma^{(1)}, ..., \gamma^{(K)} \in \mathbf{R}_0^+$ of G. Find the optimal parameter $\lambda_0 = \mathrm{argmax}_\lambda \mathcal{L}(\lambda)$ (as a function of $\gamma^{(1)}, ..., \gamma^{(K)}$) that maximizes the empirical log-likelihood

$$\mathcal{L}(\lambda) = \langle \ln q_G \rangle \approx K^{-1} \sum_{k=1}^{K} \ln q_G(\gamma^{(k)}) \ .$$

Hint: Look for zero-crossings of the derivative $\mathrm{d}\mathcal{L}/\mathrm{d}\lambda$*!*

(b) We now introduce an L^2 regularisation and consider the regularized likelihood

$$\mathcal{L}_R(\lambda) = \mathcal{L}(\lambda) - \frac{C}{2} \, \lambda^2 \ .$$

Give an intuitive explanation of the term $\frac{C}{2}\,\lambda^2$.

(c) Find the maximum likelihood solution λ^* of the the regularized problem.

(d) Use the Taylor expansion to find the polynomial dependence of λ^* on C for small values of C. Explain the sign of the correction term.

(e) Determine the scaling behavior of λ^* for $C \to \infty$ and use this knowledge together with your result from d to sketch λ^* as a function of C.

4. (Gaussian Mixture Model)

A Gaussian mixture model (GMM) is a statistical method to estimate a probability density p via likelihood maximization, assuming p is a sum of M Gaussians (components).

(a) Load the data sets `d100.txt` and `d1000.txt` containing 1000 and 1000 realizations of a 2-d random process, respectively and employ `sklearn.mixture.GaussianMixture` to fit the model assuming $M = 2, .., 5$ components.

(b) For a data set $x_1, ..., x_N$ and model class y, the Bayesian information criterion is defined as

$$\mathrm{BIC} = -\ln p(x|\hat{\vartheta}, y)\, p(\hat{\vartheta}, y) + \frac{\#|y|}{2} \ln N \approx -\ln p(y|x)$$

where $\#|y|$ denotes the number of parameters in the model class and $\hat{\theta}$ is the parameter vector that maximizes the likelihood. The BIC is a Method implemented in `sklearn.mixture.GaussianMixture`. Compute the BIC for different M. What do you see?

(c) Interpret the BIC formula in terms of a constrained optimization/regularization problem!

(d) The GMM is technically fitted using the expectation-maximization (EM) algorithm. Prepare a short (10 minute) talk explaining this algorithm based on a text book (e.g., C. Bishop *Pattern Recognition and Machine learning*).