Exercises

Advanced Machine Learning

Fall 2019

Series 2. October 10, 2019 (Bayesian Linear Regression)

Machine Learning Laboratory

Dept. of Computer Science, ETH Zürich

Prof. Joachim M. Buhmann

Web https://ml2.inf.ethz.ch/courses/aml/

Teaching assistant: **Stefan Stark**

starks@inf.ethz.ch

Problem 1 (MLE for Gaussians):

Consider a data set $X = \{x_1, ..., x_N\}$ drawn i.i.d. from $N(\mu, \mathbf{\Sigma})$.

a. Write down the log-likelihood function of the data. [1]

b. Derive $\hat{\mu}$ and $\hat{\Sigma}$, the MLE estimates of μ and Σ. [2]

c. Show that $\hat{\mu}$ is an unbiased estimator and $\hat{\Sigma}$ is a biased estimator.

Problem 2 (Conditioning a Gaussian):

Consider a D-dimensional vector $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, partioned into

$$\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$$

with the corresponding partionings

$$\mu = (\mu_a, \mu_b)$$

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

The conditional $p(\mathbf{x}_a \,|\, \mathbf{x}_b)$ is a Gaussian distribution, $\mathcal{N}(\mathbf{x}_a \,|\, \mu_{a|b}, \Sigma_{a|b})$. Here we will derive expressions for its mean $\mu_{a|b}$ and variance, $\Sigma_{a|b}$.

1. The exponential in $\mathcal{N}(\mathbf{x} \,|\, \mu, \Sigma)$ is $-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$. Show that this can be represented by a term quadratic in $\mathbf{x}$, a term linear in $\mathbf{x}$ and a constant term that does not depend on $\mathbf{x}$, $a\mathbf{x}^T \mathbf{A} \mathbf{x} + b\mathbf{x}^T \mathbf{b} + c$. This is the multi-variate version of completing the square. Derive expressions for $\mathbf{A}$ and $\mathbf{b}$ as functions of μ and Σ.
2. Expand the exponential from part (1) in terms of the components $\mathbf{x}_a$ and $\mathbf{x}_b$.
3. The conditional distribution, $p(\mathbf{x}_a \,|\, \mathbf{x}_b)$ is realized by treating $\mathbf{x}_b$ as constant and renormalizing the joint distribution. Using the expansion derived in (2) complete the square and give expressions for $\mu_{a|b}$ and $\Sigma_{a|b}$
4. The precision matrix Λ was used here for conveinence. Usually we only have access to the covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$. Given the identities $\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$and $\Lambda_{ab} = -\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}$, show that $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \mu_b)$ and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$.

Problem 3 (Bayesian Regression):

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with n observations and p predictor variables. We model $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ and $\beta \sim N(0, \mathbf{\Lambda}^{-1})$.

---

[1] You may find the trace trick useful: $x^T A x = \mathrm{Tr}(x^T A x) = \mathrm{Tr}(x x^T A) = \mathrm{Tr}(A x x^T)$

[2] Useful calculus identities: $\frac{\partial}{\partial A}|A| = A^{-T}$, $\frac{\partial}{\partial A}\mathrm{Tr}(AB) = B^T$

1. What is the dimensionality of ϵ? Of $\mathbf{X}$? Of β?
2. Show that the posterior distribution $p(\beta|\mathbf{Y}, \mathbf{X}, \sigma^2, \mathbf{\Lambda})$ is normal with mean $\mu_\beta = (\mathbf{X^T X} + \sigma^2 \mathbf{\Lambda})^{-1} \mathbf{X^T Y}$ and covariance matrix $\mathbf{\Sigma}_\beta = \sigma^2 (\mathbf{X^T X} + \sigma^2 \mathbf{\Lambda})^{-1}$.
3. What is the dimensionality of $(\mathbf{X^T X} + \sigma^2 \mathbf{\Lambda})^{-1}$? Of μ_β? Of $\mathbf{\Sigma}_\beta$?
4. Assume the form $\mathbf{\Lambda} = \frac{\lambda}{\sigma^2}\mathbf{I}$. What affect does varying λ have on the posterior distribution of β?

Problem 4 (Intro to Prediction in Gaussian Processes):

Let $\mathbf{f}$ be the noise-free latent function value from a Gaussian Process with mean 0 and kernel K, evaluated at locations $\mathbf{X}$, i.e.

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid 0, K(\mathbf{X}, \mathbf{X}))$$

Derive the mean and variance of $\mathbf{f}_*$, latent function values evaluated at a set of new locations $\mathbf{X}_*$, conditioned on $\mathbf{f}$.