Exercises
Advanced Machine Learning
Fall 2020

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Series October 5, 2020 (Bayesian Linear Regression)

Teaching assistant: **João Carvalho**
joao.carvalho@inf.ethz.ch

Solution 1 (MLE for Gaussians):

1. Write down the log-likelihood function of the data.

$$
\begin{aligned}
p(\mathbf{X} \mid \mu, \Sigma) &= \prod_i^N \mathcal{N}(\mathbf{x}_i \mid \mu, \Sigma) \\
\mathcal{L}(\mu, \Sigma) &= \log p(\mathbf{X} \mid \mu, \Sigma) \\
&= \sum_i^N \log \mathcal{N}(\mathbf{x}_i \mid \mu, \Sigma) \\
&= \sum_i^N -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \\
&\propto N \log|\Sigma| + \sum_i^N (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) + const
\end{aligned}
$$

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

   a) Derive $\hat{\mu}$

   $\frac{\partial}{\partial \mu}\mathcal{L}(\mu, \Sigma) \coloneqq 0 \implies$

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \sum_i^N (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) &= 0 \\
\frac{\partial}{\partial \mu} \sum_i^N \mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i - 2\mu^T \Sigma^{-1} \mathbf{x}_i + \mu^T \Sigma^{-1} \mu &= 0 \\
\sum_i^N -2\Sigma^{-1}\mathbf{x}_i + 2\Sigma^{-1}\mu &= 0 \\
\frac{1}{N}\sum_i^N \mathbf{x}_i &= \hat{\mu}
\end{aligned}
$$

   b) Derive $\hat{\Sigma}$.
   We need a few matrix identities and derivatives. Namely $|A^{-1}| = |A|^{-1}$, some trace identities $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, for scalar c: $\mathrm{Tr}(c) = c$, $\frac{\partial |A|}{\partial A} = |A| A^{-T}$ and $\frac{\partial \mathrm{Tr}(A^T B)}{\partial A} = B$. A good resource

for matrix identities is the matrix cookbook:
`https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.

The derivation is simpler using $\Lambda = \Sigma^{-1}$, which is valid since Λ has a one-to-one correspondence with Σ.

$$\mathcal{L}(\mu, \Lambda) \propto -N \log|\Lambda| + \sum_{i}^{N} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu) + const$$

$$\begin{aligned}\frac{\partial}{\partial \Lambda} \log|\Lambda| &= \frac{1}{|\Lambda|} \frac{\partial |\Lambda|}{\partial \Lambda} \\ &= \frac{1}{|\Lambda|} |\Lambda| \Lambda^{-T} \\ &= \Lambda^{-T}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \Lambda} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu) &= \frac{\partial}{\partial \Lambda} \mathrm{Tr}\left((\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu)\right) \\ &= \frac{\partial}{\partial \Lambda} \mathrm{Tr}\left((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \Lambda\right) \\ &= (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T\end{aligned}$$

$$\frac{\partial}{\partial \Lambda} \mathcal{L}(\mu, \Lambda) \coloneqq 0 \implies$$

$$\begin{aligned}-N\Lambda^{-T} + \sum_{i}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T &= 0 \\ \frac{1}{N} \sum_{i}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T &= \hat{\Sigma}\end{aligned}$$

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

$$\mathbb{E}[\hat{\mu}] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\mathbf{x}_i] = \mu$$

Reminder:

$$\Sigma = \mathbb{E}[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T] = \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] - \mu\mu^T$$

and for $i \neq j$:

$$\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] = \mathbb{E}[\mathbf{x}_i]\mathbb{E}[\mathbf{x}_j^T] = \mu\mu^T$$

Therefore (δ_{ij} is the dirac delta)

$$\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] = \delta_{ij}\Sigma + \mu\mu^T$$

$$
\begin{aligned}
\mathbb{E}[\hat{\Sigma}] &= \frac{1}{N}\sum_i^N \mathbb{E}[(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T] \\
&= \frac{1}{N}\sum_i^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] - \frac{1}{N}\sum_i^N \mathbb{E}[\mathbf{x}_i\hat{\mu}^T] - \frac{1}{N}\sum_i^N \mathbb{E}[\hat{\mu}\mathbf{x}_i^T] + \mathbb{E}[\hat{\mu}\hat{\mu}^T] \\
&= \frac{1}{N}\sum_i^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] - \frac{1}{N^2}\sum_i^N\sum_j^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^T] - \frac{1}{N^2}\sum_i^N\sum_j^N \mathbb{E}[\mathbf{x}_j\mathbf{x}_i^T] + \frac{1}{N^2}\sum_i^N\sum_j^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^T] \\
&= \frac{1}{N}\sum_i^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] - \frac{1}{N^2}\sum_i^N\sum_j^N \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^T] \\
&= \frac{1}{N}N(\Sigma + \mu\mu^T) - \frac{1}{N^2}(N^2\mu\mu^T + N\Sigma) \\
&= \Sigma - \frac{1}{N}\Sigma \\
\mathbb{E}[\hat{\Sigma}] &= \frac{N-1}{N}\Sigma
\end{aligned}
$$

Solution 2 (Conditioning a Gaussian):

1. Derive expressions for μ and Σ as functions of $\mathbf{A}$ and $\mathbf{b}$.

$$
\begin{aligned}
\log p(\mathbf{x}) &\propto -\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) + const \\
&\propto -\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{x}^T\Sigma^{-1}\mu + const
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{A} = \Sigma^{-1} &\implies \Sigma = \mathbf{A}^{-1} \\
\mathbf{b} = \Sigma^{-1}\mu &\implies \mu = \Sigma\mathbf{b}
\end{aligned}
$$

2. Using the precision matrix Λ and treating $\mathbf{x}_b$ as constant, enumearte all terms in $\log p(\mathbf{x}_a, \mathbf{x}_b)$ that contain $\mathbf{x}_a$ and give expressions for $\mu_{a|b}$ and $\Sigma_{a|b}$ by completing the square.

$$
\begin{aligned}
\log p(\mathbf{x}) &\propto -\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) + const \\
&\propto -\frac{1}{2}(\mathbf{x}_a-\mu_a)^T\Lambda_{aa}(\mathbf{x}_a-\mu_a) - \frac{1}{2}(\mathbf{x}_a-\mu_a)^T\Lambda_{ab}(\mathbf{x}_b-\mu_b) \\
&\quad - \frac{1}{2}(\mathbf{x}_b-\mu_b)^T\Lambda_{ba}(\mathbf{x}_a-\mu_a) - \frac{1}{2}(\mathbf{x}_b-\mu_b)^T\Lambda_{bb}(\mathbf{x}_b-\mu_b) + const \\
\log p(\mathbf{x}_a \mid \mathbf{x}_b) &\propto -\frac{1}{2}\mathbf{x}_a^T\Lambda_{aa}\mathbf{x}_a + \mathbf{x}_a^T\left(\Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b-\mu_b)\right) + const
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{A} = \Lambda_{aa} &\implies \Sigma_{a|b} = \Lambda_{aa}^{-1} \\
\mathbf{b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b-\mu_b) &\implies \mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\mu_b)
\end{aligned}
$$

3. (Self evident)

See Bishop 2.3.1 for more details regarding this exercise.

Solution 3 (Bayesian Regression):

1. What is the dimensionality of ϵ? Of $\mathbf{X}$? Of β?
   $\epsilon : n \times 1$, $\mathbf{X} : n \times p$, $\beta : p \times 1$.
2. Derive posterior of β

$$
\begin{aligned}
p(\mathbf{y} \mid \beta) &= \mathcal{N}\left(\mathbf{X}\beta, \sigma^2 \mathbb{I}\right) \\
p(\beta) &= \mathcal{N}\left(0, \Lambda^{-1}\right)
\end{aligned}
$$

   We will complete the square, as was done in problem 2.

$$
\begin{aligned}
\log p(\mathbf{y}, \beta) &\propto -\frac{1}{2}\sigma^{-2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) - \frac{1}{2}\beta^T \Lambda \beta + const \\
\log p(\beta \mid \mathbf{y}) &\propto -\frac{1}{2}\beta^T\left(\sigma^{-2}\mathbf{X}^T\mathbf{X} + \Lambda\right)\beta + \beta^T\left(\sigma^{-2}\mathbf{X}^T\mathbf{y}\right) + const
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_\beta &= \sigma^2\left(\mathbf{X}^T\mathbf{X} + \sigma^2\Lambda\right)^{-1} \\
\mu_\beta &= \sigma^{-2}\Sigma_\beta \mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X} + \sigma^2\Lambda\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

3. $(\mathbf{X}^T\mathbf{X} + \sigma^2\Lambda)^{-1} : p \times p$, $\mu_\beta : p \times 1$, $\Sigma : p \times p$.
4. Increasing λ increases the regularization strength

Solution 4 (Prediction in Gaussian Processes):

The joint distriubtion is gaussian,

$$
p(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(0, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)
$$

Where $K = k(\mathbf{x}, \mathbf{x})$, the kernel evaulated on the observed inputs. Similarly, $K_* = k(\mathbf{x}, \mathbf{x}_*)$ and $K_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$. The predictive distribution is found by simply conditioning on the observed data.

$$
\begin{aligned}
p(\mathbf{f}_* \mid \mathbf{f}) &= \mathcal{N}(\mathbf{f}_* \mid \mu_*, \Sigma_*) \\
\mu_* &= K_*^T K^{-1}\mathbf{f} \\
\Sigma_* &= K_{**} - K_*^T K^{-1} K_*
\end{aligned}
$$