Exercises
Machine Learning
HS 2016

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

Series 2, Oct 24th, 2016 (MLE)

Email questions 1,2 to: **Andrey Ignatov** ihnatova@student.ethz.ch
Email questions 3,4,5 to: **David Zhao** zhaoda@student.ethz.ch

Problem 1 (MLE):

Suppose we have a family of probability distributions $p(x|\boldsymbol{\theta})$ that depends on parameters $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_n$, and the set $X = \{x_i\}_{i=1}^N$ generated according to some concrete $\boldsymbol{\theta}$. Our goal is to find this $\boldsymbol{\theta}$, and to do this we are looking for $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood of X:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(X|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta}). \tag{1}$$

If L is continuous and differentiable, its derivative in a stationary point is zero, and $\widehat{\boldsymbol{\theta}}$ can be found from equation: $\frac{\partial}{\partial \boldsymbol{\theta}} L(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0$. Otherwise we have to apply some optimization techniques or try to find the solution analytically. In this exercise, you have to derive $\widehat{\boldsymbol{\theta}}$ for the following distributions:

- **Exponential(λ):** $p(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}, \quad x > 0, \quad \lambda > 0, \quad \boldsymbol{\theta} = \lambda$
  Hint: it is easier to maximize the logarithm of L.

  Solution. The Likelihood of the dataset X is given by:

$$L(X|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) = \lambda^N \mathrm{e}^{-\lambda \sum_{i=1}^N x_i}$$

$$\log L(X|\lambda) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

$$\frac{\partial}{\partial \lambda} \log L(X|\lambda) = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \qquad \rightarrow \qquad \boxed{\widehat{\lambda} = \frac{N}{\sum_{i=1}^N x_i}}$$

  Since all x_i ¿ 0, the obtained $\widehat{\lambda}$ is also positive. Finally, we have to verify that $\widehat{\lambda}$ is a maximum, and for this purpose we will show that the second derivative $\frac{\partial^2}{\partial \lambda^2} \log L(X|\lambda)$ is negative:

$$\frac{\partial^2}{\partial \lambda^2} \log L(X|\lambda) = -\frac{N}{\lambda^2} < 0.$$

- **Uniform[0, θ]:** $p(x|\theta) = \begin{cases} \theta^{-1}, & x \in [0,\theta] \\ 0, & \text{else} \end{cases}, \qquad \boldsymbol{\theta} = \theta$

  Hint: $p(x|\theta)$ can be written as $p(x|\theta) = \theta^{-1} \mathcal{I}_{[0,\theta]}(x)$, where $\mathcal{I}_{[0,\theta]}(x) = \begin{cases} 1, & x \in [0,\theta] \\ 0, & \text{else} \end{cases}$.

Solution. The Likelihood of the dataset X is given by:

$$L(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \frac{1}{\theta^N}\mathcal{I}_{[0,\theta]}(x_1)\mathcal{I}_{[0,\theta]}(x_2)\dots\mathcal{I}_{[0,\theta]}(x_n)$$

Let $\theta^* = x^* = \max\{x_i\}$.

1) If $\theta < \theta^*$, then $\mathcal{I}_{[0,\theta]}(x^*) = 0 \to L(X|\theta) = 0$.

2) If $\theta > \theta^*$, then $0 < L(X|\theta) = \frac{1}{\theta^N} < \frac{1}{\theta^{*N}} = L(X|\theta^*)$.

Therefore, $\boxed{\widehat{\theta} = \theta^* = \max\{x_i\}}$

- **Bernoulli(θ):** $x \in \{0,1\}$, $\begin{cases} p(x=0|\theta) = 1-\theta \\ p(x=1|\theta) = \theta \end{cases}$ or $p(x|\theta) = \theta^x(1-\theta)^{1-x}$, $\boldsymbol{\theta = \theta}$

  Note that this is a discrete distribution.

Solution. The Likelihood of the dataset X is given by:

$$L(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \theta^{\sum_{i=1}^N x_i}(1-\theta)^{\sum_{i=1}^N 1-x_i} = \theta^S(1-\theta)^{N-S}, \qquad S = \sum_{i=1}^{N} x_i$$

$$\log L(X|\theta) = S\log\theta + (N-S)\log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\log L(X|\theta) = \frac{S}{\theta} - \frac{N-S}{1-\theta} = 0 \qquad \to \qquad \boxed{\widehat{\theta} = \frac{S}{N} = \frac{\sum_{i=1}^N x_i}{N}}$$

$$\frac{\partial^2}{\partial\lambda^2}\log L(X|\theta) = -\frac{S}{\theta^2} - \frac{N-S}{(1-\theta)^2} < 0.$$

- **Multinomial(p, m):** $\mathbf{x} = (x_1, x_2, ..., x_k)$, $\mathbf{p} = (p_1, p_2, \dots, p_k)$, $\sum\limits_{j=1}^{k} x_j = m$, $\sum\limits_{j=1}^{k} p_j = 1$, $x_j \in \overline{0...n}$

$$p(\mathbf{x}|\mathbf{p}) = \tfrac{m!}{x_1!x_2!\cdots x_k!}\, p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}, \qquad \boldsymbol{\theta = \mathbf{p}}.$$

Hint: you need to use the method of Lagrange multipliers here. If you don't know what is it, you can skip this problem now and return to it after the lecture on SVMs.

Solution. The Likelihood of the dataset X is given by:

$$L(X|\mathbf{p}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{p}) = \prod_{i=1}^{N} \frac{m!}{x_{i1}!x_{i2}!\cdots x_{ik}!}\, p_1^{x_{i1}}p_2^{x_{i2}}\cdots p_k^{x_{ik}}$$

$$\log L(X|\mathbf{p}) = \sum_{i=1}^{N}\sum_{j=1}^{k} x_{ij}\log p_j + C(X,n)$$

Since we have an additional constraint $\sum\limits_{j=1}^{k} p_j = 1$, we have to use the method of Lagrange multiplies.

The Lagrangian for this problem will take the form:

$$\mathcal{L}(X, \mathbf{p}, \lambda) = \sum_{i=1}^{N} \sum_{j=1}^{k} x_{ij} \log p_j + \lambda(\sum_{j=1}^{k} p_j - 1)$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(X, \mathbf{p}, \lambda) = \sum_{j=1}^{k} p_j - 1 = 0 \qquad \rightarrow \qquad \sum_{j=1}^{k} p_j = 1$$

$$\frac{\partial}{\partial p_j} \mathcal{L}(X, \mathbf{p}, \lambda) = \frac{\sum_{i=1}^{N} x_{ij}}{p_j} + \lambda = 0 \qquad \rightarrow \qquad \sum_{i=1}^{N} x_{ij} = -\lambda p_j$$

Lets sum both sides of the last equation over j. Since $\sum\limits_{j=1}^{k} x_{ij} = m$, and $\sum\limits_{j=1}^{k} p_j = 1$:

$$Nm = -\lambda \quad \rightarrow \quad \boxed{p_j = \frac{\sum_{i=1}^{N} x_{ij}}{-\lambda} = \frac{\sum_{i=1}^{N} x_{ij}}{Nm}}$$

Problem 2 (Conjugate Priors):

In general, the Bayesian Learning and MAP are more computationally expensive than MLE, since we have to integrate over θ in these methods:

$$p(x|X) = \int_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta}) p(\boldsymbol{\theta}|X) d\boldsymbol{\theta} \tag{2}$$

$$p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(X)} = \frac{p(\boldsymbol{\theta}) \prod p(x_i|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) \prod p(x_i|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{3}$$

The integration in 3 can be avoided if we can identify the type of the distribution $p(\boldsymbol{\theta}|X) \sim p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$, and then the normalization constant $\alpha = \int_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ can be calculated using its known parameters. In particular, if $p(\boldsymbol{\theta}|X)$ is from the same distribution family as $p(\boldsymbol{\theta})$, then the prior distribution $p(\boldsymbol{\theta})$ is called conjugate to $p(X|\boldsymbol{\theta})$. For example, the normal distribution is conjugate to the normal one.

In this exercise, you have to show that the gamma-distribution $p_{\alpha,\beta}(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \mathrm{e}^{-\beta\lambda}$ is conjugate to the exponential distribution $p(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}$.

Solution. Suppose $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$, $X \sim \mathrm{Exponential}(\lambda)$, then:

$$p(\lambda|X) \sim p(X|\lambda)\, p_{\alpha,\beta}(\lambda) = \prod_{i=1}^{N} p(x_i|\lambda)\, p_{\alpha,\beta}(\lambda) = \lambda^N \mathrm{e}^{-\lambda \sum_{i=1}^{N} x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \mathrm{e}^{-\beta\lambda} \sim \lambda^{(N+\alpha)-1} \mathrm{e}^{-\lambda(\sum_{i=1}^{N} x_i + \beta)},$$

Thus, $p(\lambda|X) \sim \mathrm{Gamma}(\tilde{\alpha}, \tilde{\beta})$, where $\tilde{\alpha} = N + \alpha$, $\tilde{\beta} = \sum\limits_{i=1}^{N} x_i + \beta$.

Problem 3 (Bayesian linear regression):

ϵ has dimension $n \times 1$, $\mathbf{X}$ has dimension $n \times p$, β has dimension $p \times 1$.

See Bayesian linear regression section of Gaussian Processes tutorial slides

$(\mathbf{X^T X} + \sigma^2 \mathbf{\Lambda})^{-1}$ has dimension $p \times p$, μ_β has dimension $p \times 1$, $\mathbf{\Sigma}_\beta$ has dimension $p \times p$

See Bayesian linear regression section of Gaussian Processes tutorial slides

Problem 4 (Kernels):

Yes, Yes, Yes, No, Yes
$k(\mathbf{x}, \mathbf{y}) = \log(\|\mathbf{x} - \mathbf{y}\| + 1)$ is not a kernel because it is not positive semi-definite. Consider two data points x, y where we let $\log(\|\mathbf{x} - \mathbf{y}\| + 1) = 1$. Then we can check that the kernel matrix $K = [[01], [10]]$ has eigenvalues $1, -1$. Because not all eigenvalues are positive, the matrix is not PSD. Alternatively, let $z = [1 - 1]^T$ and we see that $z^T K z = -2 < 0$.

Problem 5 (Gaussian Processes):

See Gaussian Processes tutorial slides.