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

Please hand in solutions until Friday Nov 4th.

("*"-exercies are a little bit more difficult, but still useful and doable)

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.
- **Uniform[0, θ]:** $p(x|\theta) = \begin{cases} \theta^{-1}, & x \in [0,\theta] \\ 0, & \text{else} \end{cases}, \quad \boldsymbol{\theta} = \theta$
  Hint: $p(x|\theta)$ can be written as $p(x|\theta) = \theta^{-1}\mathcal{I}_{[0,\theta]}(x)$, where $\mathcal{I}_{[0,\theta]}(x) = \begin{cases} 1, & x \in [0,\theta] \\ 0, & \text{else} \end{cases}$.
- **Bernoulli(θ):** $x \in \{0,1\}, \quad \begin{cases} p(x=0|\theta) = 1-\theta \\ p(x=1|\theta) = \theta \end{cases}$ or $p(x|\theta) = \theta^x(1-\theta)^{1-x}, \quad \boldsymbol{\theta} = \theta$
  Note that this is a discrete distribution.
- **Multinomial(p, m):** $\mathbf{x} = (x_1, x_2, ..., x_k), \quad \mathbf{p} = (p_1, p_2, \ldots, p_k), \quad \sum_{j=1}^{k} x_j = m, \quad \sum_{j=1}^{k} p_j = 1, \quad x_j \in \overline{0...n}$
  $$p(\mathbf{x}|\mathbf{p}) = \tfrac{m!}{x_1! x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \qquad \boldsymbol{\theta} = \mathbf{p}.$$

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.

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 $p(x|\lambda) = \lambda\mathrm{e}^{-\lambda x}$.

Problem 3 (Bayesian linear 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 $\beta \sim N(0, \mathbf{\Lambda}^{-1})$.

What is the dimensionality of ϵ? Of $\mathbf{X}$? Of β?

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 $\mathbf{\Sigma}_\beta = \sigma^2(\mathbf{X^T X} + \sigma^2\mathbf{\Lambda})^{-1}$.

What is the dimensionality of $(\mathbf{X^T X} + \sigma^2\mathbf{\Lambda})^{-1}$? Of μ_β? Of $\mathbf{\Sigma}_\beta$?

Show that Ridge is just a special case of Bayesian linear regression where $\mathbf{\Lambda} = \frac{\lambda}{\sigma^2}\mathbf{I}$.

Problem 4 (Kernels):

Which of the following are kernels? If it's not a kernel, why not?
$k(\mathbf{x}, \mathbf{y}) = C$, $C > 0$
$k(\mathbf{x}, \mathbf{y}) = \tanh(a\mathbf{x^T y}) - b$
$k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x^T y})^p$, $p \in \mathbb{N}$
$k(\mathbf{x}, \mathbf{y}) = \log(\|\mathbf{x} - \mathbf{y}\| + 1)$
$k(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{h^2})$

Problem 5 (Gaussian Processes):

Write down the equation describing a general Gaussian process. What kind of distribution does $\mathbf{Y}$ have? How do you know? Write down the distribution of $\mathbf{Y}$ and explicitly determine its parameters.

What kind of kernel function do we have if $\mathbf{\Lambda} = \lambda\mathbf{I_p}$?

What happens when we add a new observation y_{n+1}? Write down the joint distribution of all the data including the new observation. Explicitly determine the mean and covariance.

Suppose $\mathbf{\Theta}$ is a set of hyperparameters that determine the parameters $\mathbf{C_n}$, which in turn directly affect the distribution of $\mathbf{Y}$. Describe how we would in general optimize the hyperparameters.