Exercises
Introduction to Machine Learning
FS 2018

Series 1, Feb 22, 2018 (Probability and Linear Algebra)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: https://las.inf.ethz.ch/teaching/introml-s18
Email questions to:
Kfir Levy, yehuda.levy@inf.ethz.ch

A sample solutions will be published on Friday, March 2nd.

Problem 1 (Linear Regression and Ridge Regression):

Let $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be the training data that you are given. As you have to predict a continuous variable, one of the simplest possible models is linear regression, i.e. to predict y as $\mathbf{w}^T\mathbf{x}$ for some parameter vector $\mathbf{w} \in \mathbb{R}^d$.[1] We thus suggest minimizing the following loss

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2. \tag{1}$$

Let us introduce the $n \times d$ matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with the $\mathbf{x}_i$ as rows, and the vector $\mathbf{y} \in \mathbb{R}^n$ consisting of the scalars y_i. Then, (1) can be equivalently re-written as

$$\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

We refer to any $\mathbf{w}^*$ that attains the above minimum as a solution to the problem.

(a) Show that if $\mathbf{X}^T\mathbf{X}$ is invertible, then there is a unique $\mathbf{w}^*$ that can be computed as $\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

(b) Show for $n < d$ that (1) does not admit a unique solution. Intuitively explain why this is the case.

(c) Consider the case $n \geq d$. Under what assumptions on $\mathbf{X}$ does (1) admit a unique solution $\mathbf{w}^*$? Give an example with $n = 3$ and $d = 2$ where these assumptions do not hold.

The *ridge regression* optimization problem with parameter $\lambda > 0$ is given by

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}_{\text{Ridge}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \left[\sum_{i=1}^{n} \left(y_i - w^T\mathbf{x}_i\right)^2 + \lambda \mathbf{w}^T\mathbf{w}\right]. \tag{2}$$

(d) Show that $\hat{R}_{\text{Ridge}}(\mathbf{w})$ is convex with regards to $\mathbf{w}$. You can use the fact that a twice differentiable function is convex if and only if its Hessian $\mathbf{H} \in \mathbb{R}^{d\times d}$ satisfies $\mathbf{w}^T\mathbf{H}\mathbf{w} \geq 0$ for all $\mathbf{w} \in \mathbb{R}^d$ (is positive semi-definite).

(e) Derive the closed form solution $\mathbf{w}^*_{\text{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda I_d\right)^{-1}\mathbf{X}^T\mathbf{y}$ to (2) where I_d denotes the identity matrix of size $d \times d$.

(f) Show that (2) admits the unique solution $\mathbf{w}^*_{\text{Ridge}}$ for any matrix $\mathbf{X}$. Show that this even holds for the cases in (b) and (c) where (1) does not admit a unique solution $\mathbf{w}^*$.

(g) What is the role of the term $\lambda\mathbf{w}^T\mathbf{w}$ in $\hat{R}_{\text{Ridge}}(\mathbf{w})$? What happens to $\mathbf{w}^*_{\text{Ridge}}$ as $\lambda \to 0$ and $\lambda \to \infty$?

[1] Without loss of generality, we assume that both $\mathbf{x}_i$ and y_i are centered, i.e. they have zero empirical mean. Hence we can neglect the otherwise necessary bias term b.

Problem 2 (Normal Random Variables):

Let X be a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\tau^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \tau^2)$. Recall that the probability density of X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\tau} e^{-(x-\mu)^2/2\tau^2}, \quad -\infty < x < \infty.$$

Furthermore, the random variable Y given $X = x$ is normally distributed with mean x and variance σ^2, i.e. $Y|_{X=x} \sim \mathcal{N}(x, \sigma^2)$.

(a) Derive the *marginal distribution* of Y, i.e. compute the density $f_Y(y)$.

(b) Use Bayes' theorem to derive the *conditional distribution* of X given $Y = y$.

Hint: For both tasks derive the density up to a constant factor and use this to identify the distribution.

Problem 3 (Bivariate Normal Random Variables):

Let X be a bivariate Normal random variable (taking on values in $\mathbb{R}^2$) with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. The density of X is then given by

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right).$$

Find the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$.