Exercises
**Learning and Intelligent Systems**
SS 2015

# Series 1, Feb 25th, 2015 (Probability and Linear Algebra)

**LAS Group, Institute for Machine Learning**
Dept. of Computer Science, ETH Zürich
**Prof. Dr. Andreas Krause**
Web: `http://las.ethz.ch/courses/lis-s15/`
**Email questions to:**
Olivier Bachem, `olivier.bachem@inf.ethz.ch`

**It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise1` containing a PDF (LaTeXor scan) to `lis2015@lists.inf.ethz.ch` until Sunday, Mar 8th 2015.**

**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 $y_i \in \mathbb{R}$. The goal in linear regression is to find parameters $\mathbf{w} \in \mathbb{R}^d$ such that $\forall i : y_i \approx \mathbf{w}^T \mathbf{x}_i$.[1] In the lecture we considered the *least-squares* optimization problem

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

and showed that under some assumptions on $D$ there exists a unique closed form solution

$$\mathbf{w}^* = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}\mathbf{y},$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a $n \times d$ matrix with the $\mathbf{x}_i$ as rows and $\mathbf{y} \in \mathbb{R}^n$ is a vector consisting of the scalars $y_i$.

(a) Show for $n < d$ that (1) does not admit a unique solution and that $\mathbf{w}^*$ is ill-defined. Explain why in such a case we cannot uniquely identify $\mathbf{w}^*$.

(b) 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

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

(c) Show that $\hat{R}_{\text{Ridge}}(\mathbf{w})$ is convex with regards to $\mathbf{w}$ for the case $d = 1$.

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

(e) 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 (a) and (b) where (1) does not admit a unique solution $\mathbf{w}^*$.

(f) 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 $\sigma^2$, i.e. $Y|_{X=x} \sim \mathcal{N}(x, \sigma^2)$.

(a) Derive the *marginal distribution* of $Y$.

(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 $Z = X_1 - X_2 = 0$.