Exercises
Introduction to Machine Learning
FS 2018

Series 1, Feb 21st, 2019 (Probability, Analysis, 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:
Mohammad Reza Karimi, mkarimi@ethz.ch

We will publish sample solutions on Friday, Mar 8th.

Problem 1 (Sampling):

Knowing the CDF of a random variable X, enables one to draw samples from that distribution.

(a) Show that if X has distribution function F, and $U \sim \mathrm{Unif}(0, 1)$ is a uniform random number in the interval $(0, 1)$, then $F^{-1}(U)$ has the same distribution as X.

In situations where the inverse of F is not easy to compute, one can use the following method (known as the *rejection method*) for generating random variables with a density f. Suppose that γ be a function such that $\gamma(x) \geq f(x)$ for all $x \in \mathbb{R}$, and

$$\int_{-\infty}^{\infty} \gamma(x)\, dx = \alpha < \infty.$$

Then, $g(x) = \gamma(x)/\alpha$ is a probability density function. Suppose we generate a random variable X by the following algorithm:

I. Generate a random variable T with density function g.

II. Generate a random variable $U \sim \mathrm{Unif}(0, 1)$, independent of T. If $U \leq f(T)/\gamma(T)$ then set $X = T$; if $U > f(T)/\gamma(T)$ then repeat steps I and II.

(b) Show that the generated random variable X has density f.

(c) Show that the number of rejections before X is generated has a Geometric distribution. Give an expression for the parameter of this distribution.

Hints For part (a) note that

$$F^{-1}(u) = \inf\{x \mid F(x) \geq u\},$$

as F is right-continuous. For part (b), you need to evaluate

$$\mathbb{P}(T \leq x \mid U \leq f(T)/\gamma(T)).$$

Problem 2 (Multivariate Normal Distribution):

Recall the following fact about characteristic functions:

Fact 1. For a random vector X in $\mathbb{R}^d$, define its characteristic function φ_X as

$$\varphi_X(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top X)], \quad \text{for all } \mathbf{t} \in \mathbb{R}^d.$$

The characteristic function completely identifies a distribution. For a multivariate Normal distribution $\mathcal{N}(\mu, \Sigma)$, one has

$$\varphi(\mathbf{t}) = \exp(i\mathbf{t}^\top \mu - \tfrac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}).$$

(a) Let $X = (X_1, \ldots, X_d)$ be a d-dimensional standard Gaussian random vector, that is, $X \sim \mathcal{N}_d(0, I)$. Define $Y = AX + \mu$, where A is a $d \times d$ matrix and $\mu \in \mathbb{R}^d$. What is the distribution of Y? If B is an $r \times d$ matrix, what is the distribution of BY?

(b) Let X be a bivariate Normal random variable (taking on values in $\mathbb{R}^2$) with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Find the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$.

(c*) For $Y \sim \mathcal{N}_d(0, I)$, we say that the random variable $V = \|Y\|^2$ has the χ^2 (chi-square) distribution with d degrees of freedom ($V \sim \chi^2(d)$). Assume that $X_1, \ldots, X_n$ are i.i.d. samples from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$. One way to estimate σ^2 from these samples is to look at the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

where $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

Prove that $\frac{(n-1)}{\sigma^2} S^2$ has a chi-square distribution with $n - 1$ degrees of freedom.

Hint: Can you write S^2 as the norm-squared of a vector? Which vector? Take care of the dimensions.

Problem 3 (Linear Regression and Ridge Regression):

Let $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\}$ 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

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

In this exercise, $\|\cdot\|$ is always the Euclidean norm. 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

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

(d) Show that $\hat{R}_{\text{ridge}}$ is convex with respect to $\mathbf{w}$. You can use the fact that a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if for any $\mathbf{x} \in \mathbb{R}^d$ its Hessian $D^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is positive semi-definite.

[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.

(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) A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called α*-strongly convex* for some $\alpha > 0$, if for any points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ one has

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

If f is twice differentiable, an equivalent condition is that for any point $\mathbf{x} \in \mathbb{R}^d$, one has

$$D^2 f(\mathbf{x}) \succeq \alpha I,$$

which means $D^2 f(\mathbf{x}) - \alpha I$ is always positive semi-definite. Prove that a strongly convex function admits a unique minimizer in $\mathbb{R}^d$. *Hint: prove that* $f(\mathbf{x}) \to \infty$ *as* $\|\mathbf{x}\| \to \infty$.

(g) 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 $\mathbf{w}^*$.

(h) What is the role of the term $\lambda\mathbf{w}^T\mathbf{w}$ in $\hat{R}_{\text{ridge}}$? What happens to $\mathbf{w}^*_{\text{ridge}}$ as $\lambda \to 0$ and $\lambda \to \infty$? You do not need to give a complete proof, only an intuitive answer suffice.

Problem 4 (Normal Random Variables):

Let X be a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\tau^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \tau^2)$. 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.