Homework 1

MLE, MAP, Model-free, Linear Regression

CMU 10-701: Introduction to Machine Learning (Spring 2018)

OUT: Jan. 29, 2018
DUE: **Feb. 12, 2018, 10:30 AM**

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3.4"). Second, write your solution independently: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.

- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submissions can be written in LaTeX. Upon submission, label each question using the template provided by Gradescope. Please refer to Piazza for detailed instruction for joining Gradescope and submitting your homework.

- **Programming**: All programming portions of the assignments should be submitted to Gradescope as well. We will not be using this for autograding, meaning you may use any language which you like to submit.

1 MLE and its Guarantees (20 pts) [Adarsh & Wenhao]

In this question, we go step by step to explore the MLE and its statistical guarantees for the following exponential family distribution:

$$P(x|\theta^*) = h(x)\exp(\theta^*\,\phi(x) - A(\theta^*)), \tag{1}$$

where $h(x), \phi(x), A(\theta)$ are known functions, so that the distribution is specified by a single unknown parameter $\theta^* \in \mathbb{R}$. Suppose we are given n i.i.d samples $X_n = \{x_1, x_2, \ldots, x_n\}$ drawn from this distribution $P(x|\theta^*)$.

(a) **[10 pts]** MLE for Exponential Families. A key learning goal is to estimate the true parameter θ^* from the observed samples. Derive the Maximum Likelihood Estimator $\widehat{\theta}_{\text{MLE}}$ for this true parameter θ^*.

(b) **[3 pts]** Estimation Error of MLE. Suppose we are given that the true parameter satisfies:

$$\theta^* = (A')^{-1}\left(\mathbb{E}_{x\sim P(x|\theta^*)}[\phi(x)]\right).$$

And moreover that:

$$|(A')^{-1}(\theta_1) - (A')^{-1}(\theta_2)| \le L|\theta_1 - \theta_2| \;\; \forall \theta_1, \theta_2.$$

Using the above two assumptions, derive the following upper bound on the error of the Maximum Likelihood Estimate:

$$|\widehat{\theta}_{\text{MLE}} - \theta^*| \le L\left|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \mathbb{E}_{x\sim P(x|\theta^*)}[\phi(x)]\right|.$$

(c) **[7 pts]** PAC Bounds for MLE. Next, we try and estimate the sample complexity i.e. number of samples n as a function of $(L, \epsilon, \sigma^2, \delta)$ required by $\widehat{\theta}_{\text{MLE}}$ to get within an ϵ-error of the true parameter θ^*:

$$|\widehat{\theta}_{\text{MLE}} - \theta^*| \le \epsilon,$$

with probability at least $1 - \delta$.

To derive this sample complexity, assume that the random variable $\phi(x)$ has a finite non-zero variance σ^2. Recall the Chebyshev's inequality: for any random variable X with mean μ and variance $\sigma^2 > 0$, we have that:

$$P(|X - \mu| > t) \le \frac{\sigma^2}{t^2}.$$

Combine Chebyshev's inequality along with result derived in previous part to derive the sample complexity.

2 A peachy MAP problem (20 pts) [Otilia and George]

Your country, Cranberryland, and its arch-nemesis, Lemonland, are competing for the title of "The world's biggest peach producer". To get ahead of the competition, Cranberryland has hired James Bond to spy on Lemonland's weekly production of peaches. Every Sunday, James Bond needs to send a radio message to Cranberryland, specifying Lemonland's peach yield for the week. The message consists of a single number, T, representing the number of tons of peaches produced by Lemonland that week. During its transmission from Lemonland to Cranberryland, the message needs to pass through K radio towers, sequentially. Lemonland heard some rumors about this scheme, so to interfere with the enemy's plan, they tinkered with the K radio towers to corrupt James Bond's message. Each radio tower now receives the message from the previous tower, adds random noise to it (for each tower $k \in \{1, ..., K\}$ the noise comes from a Gaussian distribution with known mean μ_k and standard deviation σ_k) independently of the noise added by the previous towers, and sends it forward to the next tower. Cranberryland receives the corrupted message (let's call it M), and has hired you to help them infer from it the *most probable* number or peaches, T.

Your job is to compute the *maximum a posteriori* (MAP) estimate of the sent number T, given the received message M and having the prior knowledge that T comes from a Gaussian distribution with mean μ_0 and standard deviation σ_0.

Hint: The overall message can be written as the sum of T and the K noise terms added by each of the K towers.

3 Fun with Linear Regression (30 pts) [Dimitris]

Assume a multiple input single output system or process where the dependence between the output $y \in \mathbb{R}$ and the inputs $x \in \mathbb{R}^p$ is **linear**:

$$y = w^T x + \epsilon = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + \epsilon \tag{2}$$

where $w \in \mathbb{R}^p$ and $\epsilon \ \sim \mathcal{N}(0, \sigma)$.

Remember that the probability density function of a **Gaussian** random variable $\epsilon \sim \mathcal{N}(\mu, \sigma)$ is given by:

$$p(\epsilon; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\epsilon-\mu)^2}, \tag{3}$$

whereas the probability density function of a **Laplacian** random variable $\epsilon \sim \mathcal{L}(\mu, b)$ is given by:

$$p(\epsilon; \mu, b) = \frac{1}{2b} e^{-\frac{|\epsilon-\mu|}{b}} \tag{4}$$

Our goal in this problem is to estimate $w \in \mathbb{R}^p$ from n i.i.d data samples $D = \{(y^{(i)}, x^{(i)})\}_{i=1}^n$.

3.1 MLE

Let us first estimate w when we have **no prior** information about it.

(a) [**5 pts**] Compute the likelihood of the data, $L(w) := \prod_{i=1}^n P(y^{(i)} | x^{(i)}, \sigma, w)$. (Hint: each $y^{(i)}$ is Gaussian; what is its mean and variance?)

(b) [**5 pts**] Compute the log-likelihood of the data, $\ell(w)$ and argue why the solution of the problem

$$\min_w \|Xw - Y\|_2^2 \tag{5}$$

yields the maximizer of the likelihood, $L(w)$. **Explicitly** define X and Y.

3.2 MAP estimator with Gaussian Prior

Now assume a zero-mean **Gaussian prior** for each w_i, $i = 1, 2, \ldots, p$. In other words, assume that $w_1, w_2, \ldots, w_p$ are independently distributed from a $\mathcal{N}(0, \tau)$ distribution.

(a) [**5 pts**] Compute the posterior distribution of w, $M(w)$.

(b) [**5 pts**] Compute the log of the posterior, $m(w)$ and argue why the solution of the problem

$$\min_w \|Xw - Y\|_2^2 + \lambda \|w\|_2^2 \tag{6}$$

yields the maximizer of the posterior, $M(w)$. **Explicitly** define X, Y and λ.

3.3 MAP estimator with Laplacian Prior

Now assume a zero-mean **Laplacian prior** for each w_i, $i = 1, 2, \ldots, p$. In other words, assume that $w_1, w_2, \ldots, w_p$ are independently distributed from a $\mathcal{L}(0, \rho)$ distribution.

(a) [**5 pts**] Compute the posterior distribution of w, $M(w)$.

(b) **[5 pts]** Compute the log of the posterior, $m(w)$ and argue why the solution of the problem

$$\min_{w} \|Xw - Y\|_2^2 + \lambda \|w\|_1 \tag{7}$$

yields the maximizer of the posterior, $M(w)$. **Explicitly** define X, Y and λ.

4 Multiple Choice Questions (10 pts) [Satya]

- More than one answers can be correct
- Please explain your choice in one or two sentences.

(a) [**2 pt**] Which of the following follows from the Bayes Rule:
A. $P(X|Y,Z) = \frac{P(Y|Z,X)\,P(Z|X)\,P(X)}{P(Z|Y)\,P(Y)}$
B. $P(X|Y,Z) = \frac{P(Y|X)\,P(Z|X)\,P(X)}{P(Y|Z)\,P(Z)}$
C. $P(X|Y,Z) = \frac{P(Y|Z)\,P(X|Z)\,P(X)}{P(Z|Y)\,P(Y)}$

(b) [**1 pt**] Empirical Risk Minimization is:
A: A Model-Free approach for supervised learning.
B: Can only be used for parametric models.
C: Is always computationally feasible since we only use a finite set of samples.

(c) [**2 pts**] Which of the following statements are FALSE?
A. As the number of data points grows to infinity, the MAP estimate approaches the MLE estimate for all possible priors.
B. The best possible prior with respect to statistical guarantees is the conjugate prior.
C. The MAP is always better than the MLE with fewer samples.
D. Maximizing either the log-likelihood or the likelihood yields the same optima.

(d) [**3 pts**] Suppose Y_1, Y_2, Y_3 denotes a random sample from an exponential distribution with density function:

$$f(y) = \begin{cases} (e^{-y/\theta})/\theta & y > 0 \\ 0 & elsewhere \end{cases}$$

Consider the following five estimators of θ :

$$\hat{\theta}_1 = Y_1,\ \hat{\theta}_2 = (Y_1 + Y_2)/2,\ \hat{\theta}_3 = (Y_1 + 2Y_2)/3,\ ,\hat{\theta}_4 = min(Y_1, Y_2, Y_3),\ \hat{\theta}_5 = (Y_1 + Y_2 + Y_3)/3$$

Which of the following options are TRUE?

A. $\hat{\theta}_1$ is unbiased
B. $\hat{\theta}_5$ is unbiased
C. $\hat{\theta}_4$ is biased
D. All the above

(e) [**2 pts**] Which of the following statements are TRUE?

A. Linear regression learns a linear function of the parameters.
B. Ridge regularized linear regression is always worse than the Lasso.
C. Increasing the regularization parameter λ in Lasso regression leads to sparser regression coefficients.
D. The squared error loss is the most suitable loss function for regression problems.

5 Programming Exercise (20 pts) [Sreena and Lam]

Note: Your code for all of the programming exercises should also be submitted to Gradescope. In particular, there is a separate 'programming assignment' in Gradescope to which you should upload your code, while visualizations and written answers should still be submitted within the primary Gradescope assignment. In your code, **please use comments to point out primary functions that compute the answers to each question.**

Note : For the entire programming exercise, please turn in your code in a single zipped folder that might contain multiple source code files.

Exploring Parameter Estimation

In this problem, we will contrast the MLE and MAP parameters of a probability distribution. Suppose we observe n iid samples X_1, …, X_n, drawn from a geometric distribution with parameter θ:

$$P(X_i = k) = (1-\theta)^k \theta.$$

Given these n samples, we then want to estimate the parameter θ via either the MLE or the MAP estimators.

5.1 Maximum Likelihood Estimation

(a) [**4 pts**] We will compute an approximation of the MLE, by just computing the maximum of the log-likelihood function over a given finite set of candidate parameters. Write a function plotMLE(X, theta) that takes as input a set of samples, and a set of candidate parameters θ, and produces a plot with the log-likelihood function $\ell(\theta)$ on the Y-axis, candidate parameters θ on the X-axis, and also mark that candidate parameter $\hat{\theta}$ from the given set of candidate parameters with the maximum log-likelihood (as the approximate MLE).

(b) [**4 pts**] Consider the following sequence of 15 samples:

$$X = (0, 21, 23, 8, 9, 2, 9, 0, 7, 8, 20, 9, 7, 4, 17)$$

Use your program to produce three plots: (a) with the first five samples $(0, 21, 23, 8, 9)$, (b) with the first ten, and (c) with all fifteen. For each of the three plots, for the set of candidate parameters use $0.01, 0.02, \ldots, 1.0$. What do you observe from the resulting plots? Does the estimate change across the three plots? If yes, what is its trend?

5.2 Maximum a Posteriori Estimation

(a) [**4 pts**] Write a function plotMAP(X,theta,alpha,beta) that that takes as input a set of samples, and a set of candidate parameters θ, a value for alpha, and a value for beta, and produces a plot with the log-posterior function $\ell(\theta)$ on the Y-axis, candidate parameters θ on the X-axis, and also mark that candidate parameter $\hat{\theta}$ from the given set of candidate parameters which has the maximum posterior density (as the approximate MAP). [Note : Use Beta distribution for prior]

(b) [**4 pts**] Redo the three plots you made in the previous part, but with the log-posterior function instead, and mark the MAP estimators. Set $\alpha = 1$, $\beta = 2$. Note that $B(1, 2) = 0.5$.

(c) [**4 pts**] Do you see any significant differences between the MLE and MAP estimates?