Homework 1

MLE, MAP, Model-free, Linear Regression

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

Authors: Adithya Raghuraman, Daniel Bird, Shubhranshu Shekhar

OUT: Sept 5, 2018

DUE: **Sept 19, 2018, 2:59 PM**

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. Submissions written in latex are preferred however handwritten submissions are also allowed, please follow the instructions below for each format:
  Latex: If you are submitting a Latex document each derivation/proof written between the begin{soln} and the end{soln} for that specific question.
  Handwritten: Submissions can be handwritten in which case please submit each solution on a seperate page. You are in charge of making sure your solutions are legible if we cannot read your solutions you will not be given credit for them.
  Upon submission, label each question using the template provided by Gradescope.

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

1 Probability Review (8 pts)

Monte is a big fan of Pokémon. You are going to help Monte figure out the expected time it will take him to complete the Pokedex (collect all Pokémon).

To simplify the problem, suppose each day Monte catches one Pokémon. The Pokémon he caught is equally likely among all $n = 251$ possible Pokémon. You should think of these as random draws with replacement; for instance, it is possible to catch the same Pokémon on two different days.

(Hints: you might consider using the following concepts: geometric random variable, linearity of expectation)

Let's break down the problem into steps

1) **[2 pts]** Suppose that instead of 251 possible Pokémon, there were only 2 species, Charmander and Pikachu, so $n = 2$. What is the expected number of days Monte needs to collect both species?

2) **[2 pts]** What is the expected number of days between catching the ith new Pokémon and the $(i-1)$th new Pokémon? Assume there are n different species of Pokémon (Hint: each day you have probability p of successfully catching the ith new Pokémon, and $1 - p$ probability of getting one of the old $i - 1$ species, think about what's the RV describing the number of such trials before getting the first success. Also do you think p changes before getting the ith new species?)

3) **[4 pts]** Derive an expression of the expected number of days needed to collect n different species of Pokémon. Also report the result you get by plugging in $n = 251$. (Hint: try using linearity of expectation; previous parts of the problem might be useful).

2 Bayesian Inference (5 pts)

In this problem we'll demonstrate how Bayes inference progressively updates your belief about the world using the following steps: 1) start with a prior belief, 2) perform an experiment, and 3) update your belief. Let's see how this works.

Suppose Your friend has a fair coin (probability of 0.5 for head), and a biased coin with probability 0.8 for head. Based on past, you believe that the coin your friend is using is the fair coin with probability f = 0.5 (so equal probability to whether he is using the fair or the biased coin).

1). **[2 pts]** The friend flipped the coin and it came up head, what is the probability that he flipped the fair coin? Round to 4 decimal places

2). **[2 pts]** Your belief about the probability of the coin being fair is now updated to the posterior you just computed. Using this as your new prior, you performed another flip, and the coin came up heads again. Compute the posterior probability of the coin being fair. Again, round to 4 decimal places

3). **[1 pts]** Suppose you keep getting heads for the coin flip, after how many more flips will your belief of the coin being fair drop to below 0.05?

3 Star Wars a MAP problem (20 pts)

A long time ago in a galaxy far away the Rebel Alliance and its arch-nemesis, the evil Empire, are trapped in eternal war. Rumor has it that the Empire are sending a fleet to attack the Rebel base, thankfully the Rebel Alliance has infiltrated the Empires forces with Bothan spies. The spies need to send a radio message to the Rebel forces, specifying the size of the fleet. The message consists of a single number, T the number of ships in the attacking fleet. During its transmission from the spies to the Rebel Alliance, the message needs to pass through K Empire controlled radio towers, sequentially. The Empire heard some rumors about this scheme, so to interfere with the Rebel's plan, they tinkered with the K radio towers to corrupt the 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. The Rebel Alliance receives the corrupted number (let's call it M), and has hired you to help them infer from it the *most probable* message, T.

Your job is to compute the *maximum a posteriori* (MAP) estimate of the sent message 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. Note that the parameters for the added noise are known and you should use them in your solution.

4 MLE and its Guarantees (13pts)

In this question, we go step by step to explore the MLE and its statistical guarantees for the exponential family distribution $P(x|\theta^*)$ defined as:

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

where $h(x), \phi(x), A(\theta)$ are known functions, so that the distribution is specified by a single unknown parameter $\theta^* \in \mathbb{R}$.

(a) **[10 pts]** MLE for Exponential Families. A key learning goal is to estimate the true parameter θ^* from the observed samples. Suppose we are given n i.i.d samples $X_n = \{x_1, x_2, \ldots, x_n\}$ drawn from the distribution $P(x|\theta^*)$, 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:

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

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

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

5 Fun with Linear Regression (30 pts)

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 $D = \{(y^{(i)}, x^{(i)})\}_{i=1}^n$.

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

5.2 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. (Hint: First compute the prior and then use bayes rule to obtain the posterior)

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

(b) **[5 pts]** Compute the log-likelihood 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 minimizer of the likelihood, $L(w)$. **Explicitly** define X, Y and λ.

5.3 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.

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

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

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

6 Programming Exercise (20 pts)

Note: Your code for all of the programming exercises should also be submitted to Autolab. While visualizations and written answers should still be submitted to Gradescope as a part of the rest of the homework. In your code, **please use comments to point out primary functions that compute the answers to each question.**

Note : For the entire programming exercise, you must turn in your code in a single .tar ball that might contain multiple source code files, this should not be compressed.

Exploring Parameter Estimation

In this problem, we will contrast the MLE and MAP parameters of a probability distribution. Suppose we observe n iid samples $X_1, \ldots, X_n$, 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.

6.1 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 100000 samples obtained from Casino Coruscant on number of trials to first win on a slot machine. The data can be found in your handout as hw1_dataset.txt
Use your program to produce three plots: (a) with the first thousand samples, (b) with the first ten-thousand, and (c) with all hundred-thousand. For each of the three plots, for the set of candidate parameters use $0.01, 0.02, \ldots, 0.99$. What do you observe from the resulting plots? Does the estimate change across the three plots? If yes, what is its trend?

6.2 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. $Beta(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\alpha > 0, \beta > 0$.]

(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?