Machine Learning in Computational Biology

21/3/2023

Assignment #1

Problem1 (10)

Let $\{\underline{x}_1, \dots, \underline{x}_N\}$ be N random vectors following a multidimensional Normal distribution. Assuming that the covariance matrix is known, derive analytically the Maximum Likelihood Estimate "$\underline{\mu}_{ML}$" for the distribution's mean.

Problem 2 (10)

Prove the formulas for the mean and the variance of the Binomial distribution.

Problem 3 (25)

Let x be a random variable following a Gaussian distribution $N(\mu, \sigma^2)$ with a known variance σ^2 but an unknown mean μ. As in the Bayesian framework, we believe that μ follows a prior distribution $N(\mu_0, \sigma_0^2)$. Given a data set of N independent observations $X = \{x_1, \dots, x_N\}$ show that:

3.1. The posterior distribution of the mean $p(\mu|X)$ ia also a Gaussian with mean $\mu_N = \frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, and variance $\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}$

Hint: In the proof, you may need to use the “complete the square» trick we learned in High School.

3.2 Consider now that x follows the distribution $x \sim N(\mu,16)$, and as Bayesians, we assume a prior for the mean $\mu \sim N(0,4)$. Use the distribution $N(7,16)$ to generate N observations for x.

a. Develop an algorithm that estimates the mean and variance of the posterior distribution $p(\mu|X)$, assuming we have available a dataset of *N= 1, 5, 10, 20, 50, 100* και *1000* observations, respectively. What do you observe as the number of observations N is increasing?

b. For every value of *N*, provide a diagram that shows the prior distribution, the distribution generating the data, and the estimated posterior distribution. Clearly label the axes of your diagrams.

Problem 4 (25)

Draw a period of the sinusoidal function $y(x)=sin(2\pi x)$ and select N samples for x uniformly distributed in the interval [0,1]. To every $y(x)$ add Gaussian noise distributed as $N(0,1)$ to generate a data set of noisy observations.

Fit to the noisy observations a polynomial model of degree M=2,3,4,5 or 9 and provide a table with the coefficients of the best least-squares fit model and the achieved RMSE. Also, provide a plot showing the function $y(x)$, the observations drawn, and the best fit model for every value of M.

Repeat the above procedure for two values of $N=10$ and $N=100$. What do you observe? Discuss your findings.

Problem 5 (30)

For the same setup as in Problem 4 above let's assume that the observations are generated as

$t = y(x) + \eta$, where $y(x)=sin(2\pi x)$ and the Gaussian noise η is distributed by $N(0, \beta^{-1})$ with known $\beta=11.1$. You are given a dataset generated in this way with $N=10$ samples randomy selected in the range $0<x<1$. Assume that you want to fit to the data a regression model $t = g(x,\boldsymbol{w}) + \eta$, where $g(x,\boldsymbol{w})$ is a $M=9$ degree polynomial with coefficients vector $\boldsymbol{w}$ following a Normal prior distribution with precision α=0.005 (Bayes approach), i.e., the prior for $\boldsymbol{w}$ is selected to be

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}.$$

Construct the predictive model

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

which allows us for every unseen randomly selected x (not in the training set) to produce a prediction and also quantify our uncertainty on the prediction. Plot $m(x)$ and $s^2(x)$ of the predictive Gaussian model for many different values of x selected randomly in the interval $0<x<1$. What do you observe? Discuss your findings.

Hint: You can use the formulas from the analysis presented in class.

Due date: Please upload to eclass your report by **Friday, 7/4/2023**. Your solutions should be complete, concise, and neatly presented. For problems that require coding you should use either python or R. Include all files in a zip file with filename:

YourLastNameFirstName_IDnumber_Assignment1.zip

Bonus (up to 30% for each problem): For questions that require code development, you may get "bonus points" if you also submit BOTH a python AND an R notebook that generates all the results and is adequately documented.

Please note that bonus parts are optional. Your class grade will not be impacted at all if you do not accumulate bonus points. However, bonus points can help you secure the higher grade if at the end of the semester your performance is between two grades (e.g., get an 8 instead of a 7.2 for the class).

Attention: By submitting your report for grading you are attesting that it represents your own work 100% and that you did not get solutions or code from anyone (including AI!). If we discover that this rule has been violated, all parties involved in providing or accepting solutions will automatically get a zero grade in this class, no questions asked!.

Be honest, say NO to cheating and plagiarism, it does not serve you in any way!