Tutorial 1 with Solution

CS 337 Artificial Intelligence & Machine Learning, Autumn 2021

August, 2021

Problem 1. A professor decides she will conduct an in-class quiz exactly on those days when at least j of the N registered students show up to class. Each student independently shows up to class with probability p if it is not raining and with probability q if it is raining outside. Suppose on a given day, the probability that it will rain is r. Provide an expression for the probability that the professor conducts an in-class quiz on that day.

Problem 2. A geometric random variable X with parameter p has the following probability mass function:

$$P(X = i) = p(1-p)^{i-1}, \qquad i = 1, 2, \ldots$$

Suppose X and Y are both independent, identically distributed, geometric random variables with parameter p. Show that

$$P(X = i | X + Y = n) = \frac{1}{n-1}, \qquad i = 1, \ldots, n-1$$

Problem 3. If $\mathbf{u}$ and $\mathbf{v}$ are two orthonormal vectors in $\mathbb{R}^8$, what is the formula for the projection of any vector $\mathbf{a}$ onto the plane spanned by the two vectors $\mathbf{u}$ and $\mathbf{v}$? How does this formula change if $\mathbf{u}$ and $\mathbf{v}$ are unit vectors but not orthogonal to each other?

Problem 4. We discussed in class that maximizing a monotonically increasing function of an objective is more convenient (and at times more intuitive) than maximizing the original objective. Prove that maximizing a strictly monotonically increasing transformation of the objective gives the same optimality point as does maximizing the original objective. (Hint: Prove by contradiction)

Answer: We will prove by contradiction. Let $O(\theta)$ be the objective function being maximized. Let $\theta^* = \underset{\theta}{\text{argmax}}\ O(\theta)$. Let $f(\beta)$ be a monotonically increasing function. Let $\hat{\theta} = \underset{\theta}{\text{argmax}}\ f(O(\theta))$ such that $\hat{\theta} \neq \theta^*$ and $f(O(\hat{\theta})) > f(O(\theta^*))$. Since f is a monotonically increasing function of its arguments, it must be that $O(\hat{\theta}) > O(\theta^*)$. Which is a contradiction, since we had $\theta^* = \underset{\theta}{\text{argmax}}\ O(\theta)$. Thus either, it must be that $\hat{\theta} = \theta^*$ OR $f(O(\hat{\theta})) = f(O(\theta^*))$.

Problem 5. A student of CS 337 measured the height of each student in class along with his/her age, weight and the heights of the child's parents. The student fit a linear regression model to predict the height as a function of the other observations i.e., height = f(age, weight, height of the child's mother, height of the child's father). The student suddenly realizes that she measured the height of each child with his/her shoe on. This meant that the estimated model had to be corrected somehow. It was known that every child in the school wore a shoe of 1.5 mm thickness. What is the simplest correction the ML student can apply to the model (without computing it all over again) to get the same result as she would have obtained by using the correct height (excluding the shoe thickness)? Justify your answer.

Solution:
Consider

1. $\hat{\mathbf{w}_{ML}} = \text{argmin} \sum_{j=1}^{m} (\mathbf{w}^T \phi(\mathbf{x}_j) + b - y_j)^2$

2. $\hat{\mathbf{w}_{ML}} = \text{argmin} \sum_{j=1}^{m} (\mathbf{w}^T \phi(\mathbf{x}_j) + b_{new} - (y_j + 1.5))^2$

Claim is $b_{new} = b - 1.5$, where one can obtain the solution b_{new} to (2) simply by subtracting 1.5 mm from b. If not, then one can show that the solution $b_{new} \neq b - 1.5$ to the second should have yielded a better solution $b_{new} + 1.5 \neq b$ to (1) - which is a contradiction.

Problem 6. Consider a random variable Y that is generated, conditional on X, based on the following process:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y = aX + \epsilon$$

Assume we have a training dataset of m pairs $\{x_i, y_i\}$ for $i = 1..m$, and σ is known. Analytically derive the correct expression for the maximum likelihood estimate of the parameter a in terms of x_i's and y_i's.

Solution:

This is a very special case of maximum likelihood estimation for linear regression with $\phi(\mathbf{x}) \in \Re$ (that is, with $p = 1$). This can be also be solved by using $\mathbf{w} = (\mathbf{\Phi^T \Phi})^{-1}\mathbf{\Phi^T y}$ and substituting $\Phi = [x_1; x_2; ...x_m]$ and $\mathbf{w} = [a]$ to get $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$

Another way of proving $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$ is from first principles. Here is the second method:

Solve for $\arg\max_a \prod_i exp\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)$. Equivalently, you could also solve for maximizing the monotonically increasing (log) transformation of the objective

$$\arg\max_a \log\left(\prod_i exp\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)\right) = \sum_i \left(\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)\right)$$

Taking partial derivative w.r.t. a we get $\sum_i ax_i^2 - x_i y_i = 0$. That is, $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$