CS 189 Introduction to Machine Learning
Spring 2018 DIS2

1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations $y_1, y_2, \ldots, y_n$ distributed according to $p_\theta(y_1, y_2, \ldots, y_n)$ (here p_θ can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as $L(\theta) = p_\theta(y_1, y_2, \ldots, y_n)$ and the MLE is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L(\theta).$$

We often make the assumption that the observations are *independent and identically distributed* or iid, in which case $p_\theta(y_1, y_2, \ldots, y_n) = f_\theta(y_1) \cdot f_\theta(y_2) \cdot \cdots \cdot f_\theta(y_n)$.

(a) Your friendly TA recommends maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ instead of $L(\theta)$. Why does this yield the same solution $\hat{\theta}_{\text{MLE}}$? Why is it easier to solve the optimization problem for $\ell(\theta)$ in the iid case? Write down both $L(\theta)$ and $\ell(\theta)$ for the Gaussian $f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y-\mu)^2}{2\sigma^2}}$ with $\theta = (\mu, \sigma)$.

(b) What is $\int_{y_1, y_2, \ldots, y_n} p_\theta(y_1, y_2, \ldots, y_n)\, dy_1 \cdots dy_n$? What is $\int_\theta p_\theta(y_1, y_2, \ldots, y_n) d\theta$?

(c) The Poisson distribution is $f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Let $Y_1, Y_2, \ldots, Y_n$ be a set of independent and identically distributed random variables with Poisson distribution with parameter λ. Find the joint distribution of $Y_1, Y_2, \ldots, Y_n$. Find the maximum likelihood estimator of λ as a function of observations $y_1, y_2, \ldots, y_n$.

2 Intuition for Regularization

This problem is meant to be a short toy problem that explores the trade-off between the complexity of a model and its fit to the training data. If we make our models complicated enough, then we can fit pretty much anything. We avoid this using regularization. Before we do the math of regularization, let's think about why it's necessary.

One of our friendly TAs frequently goes to the pool to prepare for her upcoming triathlon. In the past month, she went to the pool 19 times. We want to look at the data and predict when she will go to the pool next month. Here is the data from this month:

Week 1: Sunday, Monday, Tuesday, Thursday, Friday
Week 2: Monday, Tuesday, Thursday, Saturday

Week 3: Sunday, Monday, Thursday, Friday, Saturday
Week 4: Monday, Tuesday, Thursday, Friday, Saturday

Consider the following four explanations for her swimming habits:

1. "She goes to the pool every day"
2. "She goes to the pool every day except Wednesdays"
3. "She goes to the pool every day except Wednesdays and even Sundays"
4. "She goes to the pool every day except Wednesdays and even Sundays and the Tuesday of third week of the month and the Saturday of the first week of the month and the Friday of the second week of the month."

(a) What is the right criterion for what makes a "good" rule? How can we validate it? Which rule is intuitively the best?

(b) Out of the 28 days in the four weeks, how many does each rule get right? That is, on how many days does the rule predict whether or not she will be at the pool?

(c) Now, we have an intuitive notion that the later rules are more complex than the earlier rules. This is a rather difficult notion to formalize, but let's give it a shot by trying to measure the number of "cases" created by each rule. If we wanted to implement these rules in code, how many "case" statements (if/elif/elif/elif/else) would it take? Let's assume that we're only allowed to use AND statements, not OR statements (so that we can't combine cases).

(d) Which rule is the "best" rule if we only measure how well it does on the training data?

(e) Which rule is the "best" rule if we require that the "complexity" (as measured by the number of cases) is 2 or less?

(f) Now, which rule is best if we don't set a strict requirement on the number of cases, but we penalize our correctness by 2 points for each case?

(g) Does this match your intuition? How does this relate to regularization?

(h) On Week 5, the schedule was "Sunday, Monday, Tuesday, Thursday, Friday, Saturday" and on Week 6 it was "Tuesday, Thursday, Friday, Saturday". How does this change your conclusions?

(i) Is there any relation between this problem and ridge regression?

3 Recalling Determinants

For a square matrix $\mathbf{A}$, the determinant $\det(\mathbf{A})$ is the oriented volume of the parallelepiped spanned by the columns or rows of $\mathbf{A}$. We then have (a) $\det(\prod_i \mathbf{A}_i) = \prod_i \det(\mathbf{A}_i)$, (b) $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$, (c) for $\mathbf{T}$ an upper triangular matrix, $\det(\mathbf{T})$ is the product of the diagonal entries of $\mathbf{T}$ and (d) the determinant of a matrix is the product of its eigenvalues.

How can you compute the determinant of an arbitrary matrix using these properties? Think about writing the matrix as a product of elementary matrices! Find the determinant of $\mathbf{A} = \begin{bmatrix} 2 & 2 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$.