Statistical Machine Learning GR5241
Spring 2023

Homework 1

Due: Thursday, February 9th at 11:59pm

Homework submission: Please submit your homework as a pdf on Canvas.

Problem 1 (Training Error and Test Error, 25 points)

This exercise is a simulation based version and extension of James 3.7.4.

I collect a set of data ($n = 100$ observations) containing a single feature and a quantitative response. I then fit a linear regression model to the data, as well as separate quadratic regression and cubic regressions, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ and $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

1. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training mean square error for the linear regression, and also the training mean square errors for the quadratic and cubic regressions. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer through simulation.
2. Answer (1) using test rather than training.
3. To further study this problem:
   (a) Consider simulating a single realized dataset ($R = 1$) and computing the training error and test error for different values of model complexity. Construct a single graphic displaying the training error and the test error as a function of model complexity (or flexibility). Would you expect the pattern to be consistent with a different simulated dataset?
   (b) Consider simulating several realized datasets ($R = 1000$) and computing the training error and test error for different values of model complexity. In this case you will average over all $R = 1000$ realized errors to estimate the true expected training and testing error. Construct a single graphic displaying the training error and the test error as a function of model complexity (or flexibility).

Problem 2 (Maximum Likelihood Estimation, 25 points)

In this problem, we analytically derive maximum likelihood estimators for the parameters of an example model distribution, the gamma distribution.

The gamma distribution is univariate (one-dimensional) and continuous. It is controlled by two parameters, the *location parameter* μ and the *shape parameter* ν. For a gamma-distributed random variable X, we write $X \sim \mathcal{G}(\mu, \nu)$. $\mathcal{G}$ is defined by the following density function:

$$p(x|\mu, \nu) := \left(\frac{\nu}{\mu}\right)^{\nu} \frac{x^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu x}{\mu}\right) ,$$

where $x \geq 0$ and $\mu, \nu > 0$.[1] Whenever $\nu > 1$, the gamma density has a single peak, much like a Gaussian. Unlike

[1] The symbol Γ denotes the distribution's namesake, the *gamma function*, defined by

$$\Gamma(\nu) := \int_0^{\infty} e^{-t} t^{\nu-1} dt .$$

The gamma function is a generalization of the factorial to the real line: $\Gamma(n) = (n-1)!$ for all $n \in \mathbf{N}$. Fortunately, we will not have to make explicit use of the integral.

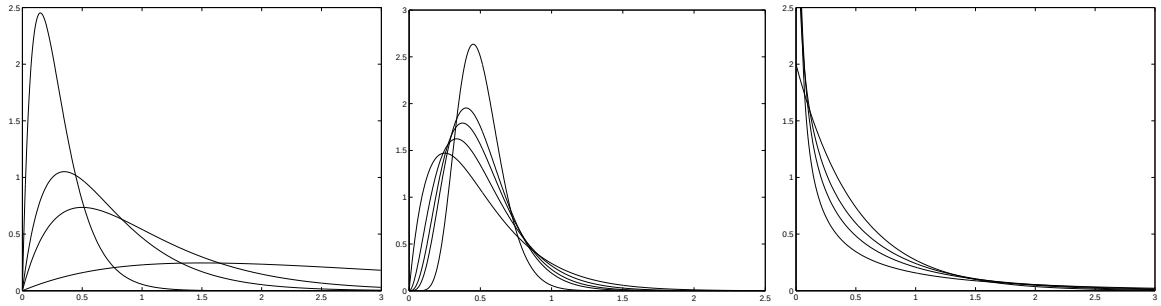

Figure 1: *Left:* The plot shows the density for different values of the location parameter ($\mu = 0.3, 0.5, 1.0, 3.0$), with the shape parameter fixed to $\nu = 2$. Since $\nu > 1$, the densities peak. As we increase μ, the peak moves to the right, and the curve flattens. *Middle:* For $\mu = 0.5$ fixed, we look at different values of the shape parameter ($\nu = 2, 3, 4, 5, 19$). Again, all the densities peak, with the peak shifting to the right as we increase ν. *Right:* If $\nu < 1$, the density turns into a monotonously decreasing function. The smaller the value of ν, the sharper the curve dips towards the origin.

the Gaussian, it is not symmetric. The first two moment statistics of the gamma distribution are given by

$$\mathsf{E}[X] = \mu \qquad \text{and} \qquad \mathsf{Var}[X] = \frac{\mu^2}{\nu} \tag{1}$$

for $X \sim \mathcal{G}(\mu, \nu)$. The plots in Figure 1 should give you a rough idea of what the gamma density may look like and how different parameter values influence its behavior.

Homework questions:

1. Write the general analytic procedure to obtain the maximum likelihood estimator (including logarithmic transformation) in the form of a short algorithm or recipe. A few words are enough, but be precise: Write all important mathematical operations as formulae. Assume that data is given as an i. i. d. sample $x_1, \ldots, x_n$. Denote the conditional density in question by $p(x|\theta)$, and the likelihood by $l(\theta)$. Make sure both symbols show up somewhere in your list, as well as a logarithm turning a product into a sum.

2. Derive the ML estimator for the location parameter μ, given data values $x_1, \ldots, x_n$. Conventionally, an estimator for a parameter is denoted by adding a hat: $\hat{\mu}$. Considering the expressions in (1) for the mean and variance of the gamma distribution, and what you know about MLE for Gaussians, the result should not come as a surprise.

3. A quick look at the gamma density will tell you that things get more complicated for the shape parameter: ν appears inside the gamma function, and both inside and outside the exponential. Thus, instead of deriving a formula of the form $\hat{\nu} := \ldots$, please show the following: Given an i. i. d. data sample $x_1, \ldots, x_n$ and the value of μ, the ML estimator $\hat{\nu}$ for the gamma distribution shape parameter solves the equation

$$\sum_{i=1}^{n} \left(\ln\left(\frac{x_i \hat{\nu}}{\mu}\right) - \left(\frac{x_i}{\mu} - 1\right) - \phi(\hat{\nu}) \right) = 0 \,.$$

   The symbol ϕ is a shorthand notation for

$$\phi(\nu) := \frac{\frac{\partial \Gamma(\nu)}{\partial \nu}}{\Gamma(\nu)} \,.$$

   In mathematics, ϕ is known as the *digamma function*.

Problem 3 (Bayes-Optimal Classifier, 25 points)

Consider a classification problem with K classes and with observations in $\mathbb{R}^d$. Now suppose we have access to the true joint density $p(\mathbf{x}, y)$ of the data $\mathbf{x}$ and the labels y. From $p(\mathbf{x}, y)$ we can derive the conditional probability $P(y|\mathbf{x})$, that is, the posterior probability of class y given observation $\mathbf{x}$.

In the lecture, we have introduced a classifier f_0 based on p, defined as

$$f_0(\mathbf{x}) := \arg\max_{y \in [K]} P(y|\mathbf{x}) \ ,$$

the *Bayes-optimal classifier*.

Homework question: Show that the Bayes-optimal classifier is the classifier which minimizes the probability of error, under all classifiers in the hypothesis class

$$\mathcal{H} := \{ f \colon \mathbb{R}^d \to [K] \mid f \text{ integrable } \} \ .$$

(If you are not familiar with the notion of an integrable function, just think of this as the set of all functions from $\mathbb{R}^d$ to the set $[K]$ of class labels.)

Hints:
- The probability of error is precisely the risk under zero-one loss.
- Try solving this exercise using $K = 3$ classes, then generalize the result to K classes.
- You can greatly simplify the problem by decomposing the risk $R(f)$ into conditional risks $R(f|\mathbf{x})$:

$$R(f|\mathbf{x}) := \sum_{y \in [K]} L^{0\text{-}1}(y, f(\mathbf{x})) P(y|\mathbf{x}) \qquad \text{and hence} \qquad R(f) = \int_{\mathbb{R}^d} R(f|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \ .$$

  If you can show that f_0 minimizes $R(f|\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$, the result for $R(f)$ follows by monotonicity of the integral.

Problem 4 (MAP Dirichlet-Multinomial Model, 25 points)

Consider observing n dice rolls $x_1, \ldots, x_n$, where $x_i \in \{1, \ldots, K\}$, i.e., n rolls from a K-sided die. Each index $\{1, \ldots, K\}$ takes on probabilities $\theta_1, \ldots, \theta_K$, such that $\sum_{j=1}^K \theta_k = 1$. Assuming the data is iid, the likelihood has the form

$$L(\theta_1, \ldots, \theta_K | x_1, \ldots, x_n) = f(x_1, \ldots, x_n | \theta_1, \ldots, \theta_K) = \prod_{k=1}^K \theta_k^{n_k},$$

where $n_k = \sum_{i=1}^n \mathbf{1}(y_i = k)$ is the number of times event k occurred. In this setting we choose a Dirichlet prior on parameters $\theta_1, \ldots, \theta_K$. The Dirichlet prior has pdf

$$q(\theta_1, \ldots, \theta_K) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^K \theta_k^{\alpha_k - 1},$$

with support $\theta_k \in (0, 1), k = 1, \ldots, K$ and $\sum_{k=1}^K \theta_k = 1$. The normalizing constant of the prior is

$$\frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}.$$

Homework questions:

1. Derive the posterior distribution
$$\Pi(\theta_1, \ldots, \theta_K | x_1, \ldots, x_n)$$
Note that the symbol "small capital pi" (Π) is the posterior pdf where the symbol "large capital pi" ($\prod_{k=1}^{K}$) is a product.

2. Derive the *"Bayesian" maximum a posterior* (**MAP**) estimator of parameters $\theta_1, \ldots, \theta_K$.

3. Derive the *"frequintist" maximum likelihood estimator* (**MLE**) of parameters $\theta_1, \ldots, \theta_K$.

Hints:
• Use the "kernel-trick" to derive the posterior. This is an easy problem.
• You should use the Lagrangian multiplier technique to solve for the **MAP** and **MLE**.