Statistical Inference and Learning- Home Exam

Yaniv Tenzer and Boaz Nadler

Due on Aug. 7, 2019,
email submissions: `boaz.nadler@weizmann.ac.il`

Q1[Ridge regression] Consider a standard linear regression setting: We observe n samples (x_i, y_i) of the form

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i,\ \epsilon_{n\times 1} \sim N(\mathbf{0}_{n\times 1}, \sigma^2 I_{n\times n}).$$

Let $\hat{B}$ be the standard OLS solution, $\hat{B} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T y}$, where $\mathbf{X}$ is the design matrix.

(a). What is the mean and variance of $\hat{B}$? Is this estimation unbiased ? Suppose that the matrix $\mathbf{X^T X}$ has few large eigenvalues and several eigenvalues rather small (namely the matrix is ill-conditioned). Would you recommend using the least squares estimate $\hat{B}$ in this case?

One approach to overcome this ill-conditioning, is to consider the following *penalized* optimization problem for some $\lambda > 0$:

$$\arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

(b) Show that the solution to the above optimization problem is $\tilde{\beta} = \left(\mathbf{X^T X} + \lambda I\right)^{-1}\mathbf{X^T y}$,

(c) Show that $\tilde{\beta}$ is a *biased* estimator of β.

(d) Denote by $V(\hat{\beta})$ a vector with the variances of the individual components of an estimator $\hat{\beta}$. Show that for all i, $V(\tilde{\beta})_i \leq V(\hat{B})_i$.

Q2 We say that a r.v. X follows a chi-square inverse distribution with ν degrees of freedom, denoted by $X \sim \chi_\nu^{-2}$, iff $1/X \sim \chi_\nu^2$.

1. Let $Y_1, \ldots, Y_m \sim N(\mu, \sigma^2)$, and assume that μ is known. Assume that the prior distribution of σ^2 is $a\chi_\nu^{-2}$, for some parameter $a > 0$. Show that the prior density function of σ^2 is given by:

$$\pi_{a,\nu}(\sigma^2) = \frac{a^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)} \left(\sigma^2\right)^{-(\frac{\nu}{2}+1)} e^{-\frac{a}{2\sigma^2}}.$$

   Is this prior distribution self-conjugate under the Gaussian setting?

2. Find the prior distribution according to Jeffrey's rule and its corresponding posterior.
3. Next we consider a general setting where the observed data consists of n independent realizations $\mathbf{z} = (z_1, \ldots, z_n)$ or a random variable $Z \sim f_Z(z, \theta)$. Let $\hat{\theta}(\mathbf{z})$ be a pointwise estimator of θ, and define its loss as $L(\hat{\theta}, \theta) \equiv \|\theta - \hat{\theta}\|_2^2$. Define the *risk* of an estimator $\hat{\theta}(\mathbf{x})$ to be $R(\hat{\theta}) \equiv \mathbb{E}_{\mathbf{Z}}[L(\hat{\theta}, \theta)]$. We think of θ as a random variable, with some unknown distribution. Let $\pi(\theta)$ be a prior distribution on θ, and define the *Bayes risk* of $\hat{\theta}$ to be $\mathbb{E}_\pi\left[R(\hat{\theta})\right]$. Show that $\theta^*_{Bayes}(\mathbf{z}) \equiv \mathbb{E}[\theta|\mathbf{z}]$ minimizes the Bayes risk. Note that θ^*_{Bayes} is known as the *Bayes estimator*.

4. Based on your answers to (1) and (2), find the Bayes estimator and the corresponding risk of each case, and compare between the two.

Q3 Let $(x_1, \ldots, x_n)$ be n i.i.d. observations from a probability distribution with density $p(x)$. Recall that in class we considered kernel density estimator of the form

$$\hat{p}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \tag{1}$$

with a suitably chosen kernel function K.

1. Suppose that $p(x)$ is a smooth density such that its second derivative is smooth and bounded, and in particular satisfies

$$|p''(x) - p''(y)| \leq L|x - y| \qquad \forall x, y \in \mathbb{R}$$

What is then an upper bound on the mean squared error $\mathbb{E}[(\hat{p}(x_0) - p(x_0))^2]$ at some fixed point x_0 and how does it depend on n? What are the conditions that the kernel K must satisfy for this upper bound to hold ?

2. In practice we need to estimate the bandwidth h. A common method is leave-one-out cross-validation. Explain this method and the resulting formula for estimating the bandwidth.

3. In some cases, we know a-priori that the density $p(x)$ has a compact support in an interval I. For example, if x is a physical quantity that cannot be negative then $x \geq 0$, and $I = [0, \infty)$. Let us study what happens to the kernel density estimate (1) for points near the boundary, when the kernel K is symmetric and supported on [-1,1].

To this end, write $x = hz$, where $z \in [0, 1]$. Show that

$$\mathbb{E}[\hat{p}(hz)] = a_0(z)p(0) - h(a_1(z) - za_0(z))p'(0) + O(h^2).$$

where $a_j(z) = \int_{-1}^{z} u^j K(u)du$.

4. In particular what is $\mathbb{E}[\hat{p}(0)]$? Is it a consistent estimator of $p(0)$ as $n \to \infty$ and $h \to 0$? Suggest a correction method to give a consistent estimate of $p(0)$.

Q4 LDA=Linear Discriminant Analysis.

Consider a binary classification problem. We have a pair of random variables (X, Y) where $X \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$, with the following explicit distribution:

$$\text{if } Y = 1 \text{ then } X \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$$
$$\text{if } Y = -1 \text{ then } X \sim N(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$$

and with $\Pr[Y = 1] = \Pr[Y = -1] = 1/2$.

A classifier is a function $f : \mathbb{R}^d \to \{-1, 1\}$. We measure the risk of a classifier by its average (generalization) error rate,

$$R(f) = \mathbb{E}_{(X,Y)}[\mathbf{1}(f(X) \neq Y)] = \Pr[f(X) \neq Y]$$

(a) Prove that the *optimal* classifier $f^* = \text{argmin}\, R(f)$ is given by

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \|\boldsymbol{x} - \boldsymbol{\mu}\| < \|\boldsymbol{x} + \boldsymbol{\mu}\| \\ -1 & \textit{otherwise} \end{cases}$$

Show that an equivalent representation is $f(\boldsymbol{x}) = sign(\boldsymbol{x}^T \boldsymbol{\mu})$. Assume $d = 2$ and $\boldsymbol{\mu} = (1, 2)$. Plot the two centers and the decision boundary.

(b) Prove that the error rate of the optimal classifier is

$$R(f^*) = \int_{-\infty}^{-\|\boldsymbol{\mu}\|/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

The quantity $\|\boldsymbol{\mu}\|/\sigma$ is the signal-to-noise ratio of this problem. Show that the error rate is exponentially small in this quantity.

(c) In practice, even if the two classes indeed follow the assumed Gaussian model, the value of $\boldsymbol{\mu}$ is typically unknown. Suppose we have a labeled data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $n/2$ samples are from class 1 and another $n/2$ samples are from class -1.

A common approach is then to estimate $\boldsymbol{\mu}$ by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum y_i \boldsymbol{x}_i$$

and construct the plug-in classifier $\hat{f}_n$ that uses $\hat{\boldsymbol{\mu}}$ instead of $\boldsymbol{\mu}$, namely $\hat{f}_n(\boldsymbol{x}) = sign(\boldsymbol{x}^T \hat{\boldsymbol{\mu}})$.

Prove that $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \sim N(0, \frac{\sigma^2}{n}\boldsymbol{I})$. What is $\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$? Prove that the probability that $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is far from its expected value is exponentially small in n. Namely that this quantity is tightly concentrated around its mean.

(d) Suppose that both $n \gg 1$ and $d \gg 1$. Show that the effective signal to noise ratio of this classifier is smaller, of the approximate form for some scalar α,

$$\frac{\|\boldsymbol{\mu}\|}{\sigma} \frac{1 - \frac{\sigma}{\|\boldsymbol{\mu}\|}\frac{\alpha}{\sqrt{n}}}{1 + \frac{\sigma}{\|\boldsymbol{\mu}\|}\frac{\alpha}{\sqrt{n}} + \frac{1}{2}\frac{\sigma^2}{\|\boldsymbol{\mu}\|^2}\frac{d}{n}}$$

(e) Simulation study: Generate labeled data from this mixture model with $n = 100$, namely 50 samples from each class, and with a vector $\boldsymbol{\mu} = (1, 1/2, 1/4, 1/8, \ldots)$, and $\sigma = 1$. For different dimensions $d = 20, 50, 100, 500, 1000, 5000, 10000, 50000$.

For each dimension d, estimate $\hat{\boldsymbol{\mu}}$, the corresponding $\hat{f}_n$ and its error rate on a test set of 10,000 independent samples. Plot a graph of this error rate versus the dimension.

Also plot on this graph a horizontal line with the Bayes error of the optimal classifier f^*.