DS102 - Discussion 3

Wednesday, 18th September, 2019

1. Suppose that $x = (x_1, \ldots, x_n)$ are fixed d-dimensional vectors, and suppose we have observations:
$$y_i = \beta^\top x_i + \epsilon_i, i \in \{1, \ldots, n\},$$
where $\epsilon_i \sim N(0, \sigma^2)$ are independent and $\beta \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}^+$ are unknown. We denote $y = (y_1, \ldots, y_n)$. This implies that the conditional distribution of y_i given x_i is Gaussian with mean $\beta^\top x_i$:
$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta^\top x_i)^2\right).$$

   (a) Show that finding the maximum likelihood estimate of β, given $x_i, y_i : i \in \{1, \ldots, n\}$, simplifies to a least squares problem:
$$\min_\beta \|y - X\beta\|_2^2,$$
where the i-th row of X is equal to x_i.

Solution: The probability of our observations is
$$\begin{aligned} p(y|x) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta^\top x_i)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\top x_i)^2\right). \end{aligned}$$
We want to find the β that maximizes this expression. Since the maximizer does not change if we take the logarithm of the above function, we can consider:
$$\ell_1(\beta, \sigma^2; x, y) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\top x_i)^2.$$
Notice that the first term does not depend on β, so removing it won't change the maximizing β. Thus we want to maximize
$$\ell_2(\beta, \sigma^2; x, y) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\top x_i)^2,$$
or equivalently to minimize
$$\ell_3(\beta, \sigma^2; x, y) = \sum_{i=1}^n (y_i - \beta^\top x_i)^2.$$
This is exactly a least squares problem. In matrix notation, this is equal to
$$\ell_3(\beta, \sigma^2; x, y) = \|y - X\beta\|_2^2,$$
where the i-th row of X is equal to x_i.

(b) Find the maximum likelihood estimate of β, denoted $\hat{\beta}_{\text{MLE}}$, in terms of the matrix X and the vector y.

Solution: We take the derivative of $\ell_3(\beta, \sigma^2; x, y)$ with respect to β and set it to 0:

$$2X^\top(y - X\hat{\beta}_{\text{MLE}}) = 0,$$

and the condition we get is called "the normal equations":

$$X^\top y = X^\top X\hat{\beta}_{\text{MLE}}.$$

We obtain $\hat{\beta}_{\text{MLE}}$ as:

$$\hat{\beta}_{\text{MLE}} = (X^\top X)^{-1}X^\top y.$$

(c) Find the maximum likelihood estimate of the variance $\hat{\sigma}^2_{\text{MLE}}$.

Solution: We take the derivative of $\ell_1(\beta, \sigma^2; x, y)$ with respect to σ^2 and set it to 0:

$$-\frac{n}{2\hat{\sigma}^2_{\text{MLE}}} + \frac{1}{2\hat{\sigma}^4_{\text{MLE}}}\sum_{i=1}^n(y_i - \hat{\beta}^\top_{MLE}x_i)^2 = 0,$$

so we get

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n(y_i - \hat{\beta}^\top_{MLE}x_i)^2.$$

In matrix notation, this is equal to:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\|y - X\hat{\beta}_{MLE}\|_2^2.$$

(d) In class we saw that for general distributions on random variables X and Y, the *conditional expectation* $\mathbb{E}[Y|X]$ minimizes the expected MSE of an estimator of Y as a function of X:

$$\arg\min_f \mathbb{E}[\|Y - f(X)\|_2^2] = \mathbb{E}[Y|X]$$

Compute the conditional expectation $\mathbb{E}[y|X]$ for the vector y and the matrix X defined in this question.

Solution: In vector and matrix notation, the problem statement from the begining of the question can be rewritten as $y = X\beta + \epsilon$, where ϵ is a vector in

$\mathbb{R}^d$. Then by linearity of expectation it follows that

$$\begin{aligned}\mathbb{E}[y|X] &= \mathbb{E}[X\beta + \epsilon|X] \\ &= X\mathbb{E}[\beta]\end{aligned}$$

If we know β, then the best estimator (under the l_2 loss) is $X\beta \approx y$. In general, we don't know β, so we'd use an estimator of β.

2. Once upon a time, there was a British lady who claimed that she could tell from the taste which had been poured into the cup first, the tea or the milk. So Fisher designed an experiment to test it.

   - Eight cups of tea were prepared.
   - In four, the tea was poured first.
   - In the other four, the milk was poured first.
   - Other features of the cups of tea (size, temperature, etc.) were held constant.
   - Cups were presented in a random order (critical). The lady tasted them, and judged.
   - She knew there were four of each type.

   The null hypothesis is that the lady has no ability to taste the difference. The test statistic is the number of correct judgements.

   (a) What is the distribution of the test statistic under the null hypothesis?

   Solution: Under the null, all possible ways of lining up the lady's judgements and the truth about the tea cups should be equally likely *because of the random order of presentation.* In that case, the reasons for the ladys judgements are unknown, but we know that they have nothing to do with the truth. The lady can make one of $\binom{8}{4} = 70$ choices, and all are equally likely under the null.

   (b) What is the probability of a false discovery? In other words, what is the probability that the lady guesses correctly under the null?

   Solution: The probability of a correct guess is $1/70 = 0.0143$.

   (c) Would rejecting the null if and only if the lady guesses correctly be a valid test under level 0.05?

Solution: Yes, because $0.0143 < 0.05$.

(d) Would rejecting the null if and only if the lady guesses correctly for at least 6 cups (meaning she permuted the remaining two) be a valid test under level 0.05?

Solution: There are $\binom{8}{2}$ possible answers that have exactly 6 correct entries. Together with the remaining answer which is completely correct that gives 29 answers for which we reject the null. However, $29/70 \approx 0.414$, so this is not a valid test under level 0.05.