MLE Tutorial

C6

Problem 1

Show that if our model is poor, the maximum likelihood classifier we derive is not the best– even among our (poor) model set– by exploring the following example.

Suppose we have two equally probable categories (i.e., $P(\omega 1) = P(\omega 2) = 0.5$). Further, we know that $p(x|\omega 1) \sim N(0, 1)$ but assume that $p(x|\omega 2) \sim N(\mu, 1)$. (That is, the parameter θ we seek by maximum-likelihood techniques is the mean of the second distribution.) Imagine, however, that the true underlying distribution is $p(x|\omega 2) \sim N(1, 10^6)$.

$p(x|\omega 1) \sim N(0, 1)$;
$p(x|\omega 2) \sim N(\mu, 1)$. ($p(x|\omega 2) \sim N(1, 10^6)$.)

- What is the value of our maximum-likelihood estimate $\hat{\mu}$ in our poor model, given a large amount of data?

$$\hat{\mu} = 1$$

- What is the decision boundary arising from this maximum-likelihood estimate in the poor model?

-> $p(x|\omega_1) \sim N(0, 1)$, $p(x|\omega^2) \sim N(1, 1)$,

$P(\omega_1) = P(\omega_2) = 0.5$,

The decision boundary is $x = 0.5$

- Ignore for the moment the maximum-likelihood approach, and derive the Bayes optimal decision boundary given the true underlying distributions: $p(x|\omega_1) \sim N(0, 1)$ and $p(x|\omega_2) \sim N(1, 10^6)$. Be careful to include all portions of the decision boundary.

-> The decision boundary is given by x that satisfies: $P(\omega_1)\, p(x|\omega_1) = P(\omega_2)\, p(x|\omega_2)$

Plug in the models and take ln of both sides:

$$-\frac{x^2}{2} = -\frac{(x-1)^2}{2\times 10^6} - \ln 10^3$$

$$x_1 = -3.7169,\ x_2 = 3.7169$$

The decision regions

for ω_1 is [-3.7169, 3.7169]

for ω_2 is [-∞, -3.7169]U[3.7169,∞]

- Now consider again classifiers based on the (poor) model assumption of $p(x|\omega 2) \sim N(\mu, 1)$. Using your result, find a new value of μ that will give lower error than the maximum-likelihood classifier.

->The decision according to the poor model is not optimal in [0.5, 3.7169]. By moving the decision toward 3.7169 we can reduce the error. Thus any $1<\mu<7.4338$ will reduce the error.

- Discuss these results, with particular attention to the role of knowledge of the underlying model.

->This example shows that the parametric form is very important for maximum-likelihood estimation. If the assumed form is far from the true underlying model, the ML estimate can give larger error than other models in the same assumed family. In order to get good results using ML estimate, one needs to find the most accurate model for the unknown underlying model based on prior knowledge, experience, or experiments on some data. If several models are available, they should be evaluated and compared using some test data.

Problem 2

- Maximum likelihood methods apply to estimates of prior probability as well. Let samples been drawn by successive independent selection of state of nature ω_i with unknown probability $P(\omega_i)$. Let

$$z_{ik} = \begin{cases} 1 & \text{if the k}^{\text{th}}\text{sample belongs to } \omega_i \\ 0 & \text{otherwize} \end{cases}$$

a) Show that

$$P(z_{i1},\ldots,z_{in} \mid P(\omega_i)) = \prod_{k=1}^{n} P(\omega_i)^{z_{ik}} (1 - P(\omega_i))^{1-z_{ik}}$$

b) Show the MLE for $P(\omega_i)$ is $\hat{P}(\omega_i) = \frac{1}{n}\sum_{i=1}^{n} z_{ik}$

Interpret your results in words.

a) Solution: $P(z_{ik}=1 \mid P(\omega_i)) = P(\omega_i),\ P(z_{ik}=0 \mid P(\omega_i)) = 1 - P(\omega_i)$

Combining them together we have

$$P(z_{ik} \mid P(\omega_i)) = P(\omega_i)^{z_{ik}} \left(1 - P(\omega_i)\right)^{1-z_{ik}}$$

$z_{i1}, \ldots, z_{in}$ are drawn independently, thus

$$P(z_{i1}, \ldots, z_{in} \mid P(\omega_i)) = \prod_{k=1}^{n} P(z_{ik} \mid \omega_i) = \prod_{k=1}^{n} P(\omega_i)^{z_{ik}} \left(1 - P(\omega_i)\right)^{1-z_{ik}}$$

b) Solution:

$$\ln P(z_{i1},\ldots,z_{in} \mid P(\omega_i)) = \sum_{k=1}^{n} \left(z_{ik} \ln P(\omega_i) + (1 - z_{ik}) \ln(1 - P(\omega_i)) \right)$$

$$\frac{\partial \ln P(z_{i1},\ldots,z_{in} \mid \omega_i)}{\partial P(\omega_i)} = \sum_{k=1}^{n} \left(z_{ik} \frac{1}{P(\omega_i)} - (1 - z_{ik}) \frac{1}{1 - P(\omega_i)} \right) = 0$$

$$(1 - P(\omega_i)) \left(\sum z_{ik} \right) - P(\omega_i) \left(\sum (1 - z_{ik}) \right) = 0$$

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{i=1}^{n} z_{ik}$$

$\hat{P}(\omega_i)$ is the fraction of samples from class i among all of the samples.

Problem 3

- Write down the training steps and a set of discriminant functions for minimum error rate classification. You need to include the details.
- First we fix the notation, let $x_{i,k}^{j}$ represent the k^{th} feature in the i^{th} example of class ω_j . Assume that we have n training samples in total and we have n_1 for ω_1 , …, and n_C for ω_C

The discriminant functions for minimum error rate classification.

$$g_j(x) = \ln p(x \mid \omega_j) + \ln P(\omega_j)$$

Here for class j, we need to estimate both $P(\omega_j)$ and $p(x \mid \omega_j)$ For $P(\omega_j)$, using maximum likelihood estimation, we have

$$P(\omega_j) = \frac{n_j}{n}$$

To estimate $p(x \mid \omega_j)$, based on the assumptions that the features are statistically independent and they are normally distributed, we have

$$p(x \mid \omega_j) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_{j,k}} e^{-\frac{(x_k - \mu_{j,k})^2}{2\sigma_{j,k}^2}}$$

Here class ω_j has 2d parameters and they are estimated according to maximum likelihood estimation

$$\mu_{j,k} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,k}^{j}$$

$$\sigma^2{}_{j,k} = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{i,k}^{j} - \mu_{j,k})^2$$

Plug-in the results to the discriminant function given above and ignore the common constant, we have

$$g_j(x) = -\left(\sum_{k=1}^{d} \left(\ln \sigma^2{}_{j,k} + \frac{(x_k - \mu_{j,k})^2}{\sigma_{j,k}^2} \right) \right) + \ln n_j$$