

Suggested solution of Assignment 2

Teaching assistant: Wilson Tam (yct@cs.cmu.edu)

Due on:

Problem 1: Pattern recognition

Consider the following decision rule for a two-category $\{\omega_1, \omega_2\}$ one-dimensional problem on feature x :

Decide ω_1 if $x > \theta$; otherwise decide ω_2 where θ is a threshold.

1. Show that the probability of error for this rule is given by:

$$P(\text{error}) = P(\omega_1) \int_{-\infty}^{\theta} p(x|\omega_1) dx + P(\omega_2) \int_{\theta}^{\infty} p(x|\omega_2) dx \quad (1)$$

Solution: Two types of errors: (1) true label is ω_1 and $x \leq \theta$, or (2) true label is ω_2 and $x > \theta$.

$$\begin{aligned} P(\text{error}) &= P(x \in \omega_1, x \leq \theta) + P(x \in \omega_2, x > \theta) \\ &= P(\omega_1)p(x \leq \theta|\omega_1) + P(\omega_2)p(x > \theta|\omega_2) \\ &= P(\omega_1) \int_{-\infty}^{\theta} p(x|\omega_1) dx + P(\omega_2) \int_{\theta}^{\infty} p(x|\omega_2) dx \end{aligned}$$

2. By differentiating, show that a necessary condition to minimize $P(\text{error})$ is that θ satisfies

$$p(\theta|\omega_1)P(\omega_1) = p(\theta|\omega_2)P(\omega_2) \quad (2)$$

Solution: Use the following facts:

$$\begin{aligned} \int_{\theta}^{\infty} p(x|\omega_2) dx &= 1 - \int_{-\infty}^{\theta} p(x|\omega_2) dx \\ \text{and } \frac{\partial}{\partial \theta} \int_{-\infty}^{\theta} p(x|\omega_i) dx &= p(\theta|\omega_i) \end{aligned}$$

Eqn 2 follows after differentiating $P(\text{error})$ wrt θ .

3. Does Eqn 2 define θ uniquely?

Solution: No, it depends on the shape of the probability density functions (pdf) $p(x|\omega_i)$. The stationary points of Eqn 2 are basically the intersection(s) of $p(x|\omega_1)P(\omega_1)$ and $p(x|\omega_2)P(\omega_2)$.

4. Give an example where a value of θ satisfying the equation actually *maximizes* the probability of error.

Solution: Consider the scenario in Figure 1. The error is the area under the whole pdf according to the definition of error in problem 1.1. The error is maximized. You need to convince yourself that fewer error is made by moving the decision boundary line around.

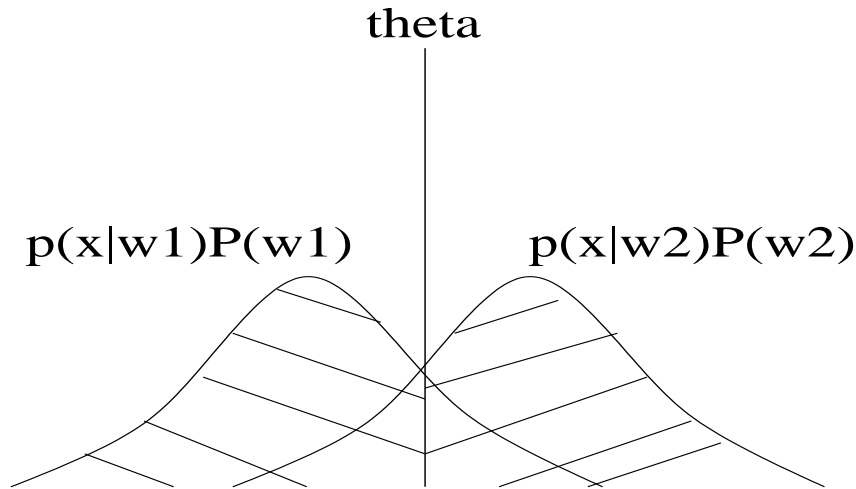


Figure 1: $P(\text{error})$ is maximized.

5. So, what is the correct decision rule which minimizes the probability of error? Explain briefly.

Solution: We should use the Bayes decision rule: Classify x as ω_1 if $p(x|\omega_1)P(\omega_1) \geq p(x|\omega_2)P(\omega_2)$, otherwise classify x as ω_2 . In other words, it is simply the rule we are familiar with: $\hat{\omega} = \underset{\omega_i}{\operatorname{argmax}} p(x|\omega_i)P(\omega_i)$. By doing so, the error is minimized as the overlapping area under the pdfs in Figure 2.

Problem 3: Acoustic modeling

1. What are the differences between discrete, continuous and semi-continuous HMMs? Compare the advantage and disadvantage of discrete HMM and continuous HMM.

Solution:

Discrete HMM: The emission probability of each state is modeled as discrete histogram over the symbols, i.e. $P(\text{symbol}|\text{state})$. The symbol is basically the index (1...K) of a codeword in a global codebook, i.e. . An input vector is quantized into a single symbol first and retrieve the probability of the symbol via a simple table lookup.

Continuous HMM: Emission probability is modeled by a continuous pdf, e.g. GMM

Semi-continuous HMM: multiple states can share the same codebook (a set of Gaussians). But the mixture weights are still state-dependent. (Some students got confused on this)

Discrete HMM: Advantage: less computation, require less data to train. Disadvantage: less accurate due to quantization error.

continuous HMM: Advantage: more accurate. Disadvantage: require more data to train, more computation

2. Suppose we want to do context clustering for the phone /t/. We already have six discrete models for that phone. They are:

T1:[1/4,3/4] with 16 training samples

T2:[1/8,7/8] with 8 training samples

T3:[3/8,5/8] with 32 training samples

T4:[5/8,3/8] with 8 training samples

T5:[7/8,1/8] with 16 training samples

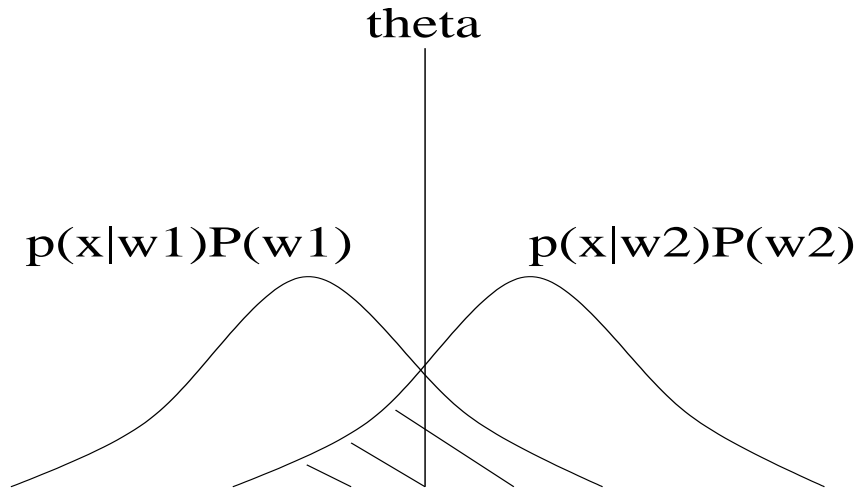


Figure 2: P(error) is minimized via Bayes decision rule.

T6:[3/4,1/4] with 4 training samples

where all models share the same Gaussians but with different distribution weights $[p_1, 1 - p_1]$ over 2 Gaussians.

Use the definition of Entropy of a discrete distribution:

$$H(p) = \sum_j p_j \log \frac{1}{p_j} \quad (3)$$

Write a program to perform the bottom-up clustering of the 6 models above into 3 clusters using the **weighted** discrete entropy distance. Attach the log file of your program as answers. [Hint: You need to consider the size of the training samples when you combine the distributions of two models.]

Solution: Many student forgot to use the weighted entropy loss weighted by the number of training samples.

Information loss after merging class i and class j is (most student got this wrong):

$$I(T_{ij}) = (n_i + n_j)H(p_{ij}) - n_iH(p_i) - n_jH(p_j) \quad (4)$$

Another point is the merging part. The number of training samples should be taken into account (most students got this right):

$$p_{ij}(x) = \frac{n_i p_i(x) + n_j p_j(x)}{n_i + n_j} \quad (5)$$