PEE
Exercise 3

Jaime Ferrando Huertas

January 2020

1 Theoretical questions

1.1 Question 1

> Briefly explain the differences between Classification and Structured Output Prediction. Cite two application examples each paradigm.

The biggest difference is the way the handle the classification output, while a classification problem outputs a label representing the class for each given sample while the Structured predicts a set of structured objects that can be mutually dependent. The space of our classification problem is finite while in Structured prediction is potentially infinite.

An example of a classification problem would be identifying the digit present on an image (the famous MNIST dataset). In this problem the output for a sample is a label between 0 and 9.

For Structured prediction an example could be trying to summarize a text. Given a sample text it should output another text that summarizes it and has a lower size. In this problem the number of outputs is potentially infinite and each individual part of the output(word) is dependent of the others in some way.

1.2 Question 2

> Justify why the naive Bayes decomposition of Eq.(5) is adequate for karyotype recognition problem

$$P(x \mid h) = P\left(x_1, \ldots, x_{22} \mid h_1, \ldots, h_{22}\right) \approx \prod_{i=1}^{22} P\left(x_i \mid h_i\right) \tag{1}$$

The naive Bayes assumption is based on the assumption of Independence between features on the samples. This is a good assumption even for features with not independent features as it helps reduce the model complexity to linear. This helps the algorithms run faster and being able to estimate a large number of parameters, translating into a powerful model.

1.3 Question 3

Briefly explain all the steps and assumptions needed to derive Eq.(9) from Eq.(7)

We will assume a deterministic feedback environment that allows us to replace a feedback recognition model with a decoding function that maps feedback signals into its decoding $d = d(f)$.

$$\widehat{h} = \arg\max_{h \in H} P\left(h \mid x, h', f\right) \tag{2}$$

We can rewrite equation 2 to:

$$\widehat{h} = \arg\max_{h \in H} \frac{P\left(h, x, h', f\right)}{P\left(x, h', f\right)} \tag{3}$$

Now we can remove the denominator since it doesn't depend on the parameter h we are trying to maximize on our argmax. This results in:

$$\widehat{h} = \arg\max_{h \in H} P\left(h, x, h', f\right) \tag{4}$$

Using the chain rule and our decoding function d we can now interpret the equation as:

$$\widehat{h} = \arg\max_{h \in H} P(d) P\left(h' \mid d\right) P\left(h \mid h', d\right) P\left(x \mid h', d, h\right) \tag{5}$$

We remove elements non dependant of h.

$$\widehat{h} = \arg\max_{h \in H} P\left(h \mid h', d\right) P\left(x \mid h', d, h\right) \tag{6}$$

And now following the slides assumption that $P(X|h', d, h)$ can be considered as independent of h' and d we obtain the resultant equation.

$$\hat{h} = \arg\max_{h \in \mathcal{H}} P(x \mid h) P\left(h \mid h', d\right) \tag{7}$$

1.4 Question 6

Briefly explain all the steps and assumptions needed to derive Eq.(19) from Eq.(7).

We will not assume feedback being deterministic so the equation will develop itself differently.

Given the same equation.

$$\widehat{h} = \arg\max_{h \in H} P\left(h \mid x, h', f\right) \tag{8}$$

We can rewrite again as

$$\widehat{h} = \arg\max_{h \in H} \frac{P\left(h, x, h', f\right)}{P\left(x, h', f\right)} \tag{9}$$

We now remove the denominator since is independent with h and include every decoding possible given f. Giving us

$$\hat{h} = \operatorname*{argmax}_{h} \sum_{d} P\left(h, x, h', d\right) \tag{10}$$

We can develop the equation given the variable dependencies.

$$\hat{h} = \operatorname*{argmax}_{h} \sum_{d} P\left(h'\right) P\left(d \mid h'\right) P(f \mid d) P\left(h \mid h', d\right) P(x \mid h) \tag{11}$$

Simplifying independent probabilities and dividing by $P(x|h)$.

$$\hat{h} = \operatorname*{argmax}_{h} P(x \mid h) \sum_{d} P\left(d \mid h'\right) P(f \mid d) P\left(h \mid h', d\right) \tag{12}$$

Given an input X and feedback f we can try to find the optimal hypothesis (h) and the optimal decoding given f.

$$(\hat{h}, \hat{d}) = \operatorname*{argmax}_{h,d} P\left(d \mid h'\right) P(f \mid d) P(x \mid h) P\left(h \mid h', d\right) \tag{13}$$

This equation is equivalent to the equation (19) in the slides.

1.5 Question 7

> Briefly explain under which conditions the solution given by Eq.(22-23) may be optimal. Do the same conditions hold for the optimality of the solution given by Eq.(20-21)? Why? Use the karyotyping example to illustrate your (otherwise general) responses

Equations 22 and 23 and be optimal in cases where the n most probable decodings have the same length as the problem size. So for the karyotyping problem the equations will hold optimal as long as n=22. This is because the n most probable decodings will contains all possible feedback to correct errors. This will make possible to calculate the optimal solution.

For equations 20-21 this will not longer be the case since they first obtain a "optimal" decoding for the feedback and then with that decoding obtain an "optimal" hypothesis. This decoding and hypothesis can join into multiple non optimal combinations and the number of those exceeds the problem size. So for n=problem size is not guaranteed to find the optimal hypothesis.

1.6 Question 8

> Briefly explain the concepts and main differences between Active and Passive interaction protocols.

The biggest difference for this protocols is the subject who decides which hypothesis element should be supervised.

In the passive protocol the expert user decides which samples needs supervision. This means the user will have to supervise all samples. There are two ways to supervise samples in this protocol

- Left-to-right, where hypothesis are supervised in a fixed order.
- Random, no specific order on hypothesis supervision process.

In the active protocol the user delegates the supervision to the system. The user only supervises a fraction of the hypothesis from the system. The system labels each element with a confident metric and ask the users to re-supervise those with the lowest scores. This system performance will be the crucial for the output quality.