STATISTICAL STRUCTURED PREDICTION
Question set 3

Ramon Ruiz Dolz

February 2019

1 Theoretical questions

1.1 (2)Question 1: Briefly explain the differences between Classification and Structured Output Prediction. Cite two application examples each paradigm.

The main difference between classification and structured prediction is how to handle the output. The Classification problem consists on given a sample, trying to get the label that represents that sample. Structured prediction are the set of problems which output variables are mutually dependent. So, while in an structured output prediction problem when giving a partial output it is important to consider all other partial outputs, in classification problems each partial output is completely independent from other outputs.

For example, the classical spam/no spam dilemma is a classification problem. To classify a determined mail into spam or no spam is not dependant of the past or next outputs of the problem, in fact determining the class only depends on the own email being classified. The spam/no spam is a non structured classification problem. An example of structured prediction problem could be the parsing problem. Given an input sequence the problem asks to build a tree whose leaves are the elements on the input and whose structure obeys some grammar. In Natural Language Processing (NLP) this problem is typified by syntactic parsing.

1.2 (2)Question 2: Justify why the naive Bayes decomposition of Eq.(5) is adequate for karyotype recognition problem.

The simplified karyotype recognition problem states the following. Given a set of 22 unsorted images of stained human chormosomes, label each image from a set of 22 labels, 1, 2, . . . 22, in such a way that each label is assigned exactly to one image. Let x be the unsorted sequence of 22 chromosomes and h be the sequence of all 22 labels.

Therefore, the naive Bayes decomposition proposed in,

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

is appropriate for the karyotype recognition problem because,

$$P(x|h) = P(x_1, \dots, x_{22}|h) = P(x_1|h)P(x_2|x_1, h) \dots P(x_{22}|x_1, x_2, \dots, x_{21}, h) \tag{2}$$

can be approximated to,

$$\approx P(x_1|h_1)P(x_2|h_2) \dots P(x_{22}|h_{22}) \tag{3}$$

since there exist independence on both x and h. The independence on the x exists because the shape of a chromosome doesn't depend on the shape of other chromosomes. Even though, to make easier the notation, the subindex of the x is in order, given the nature of this problem it is not strictly necessary that x_2 comes after x_1, or x_{15} before x_{16}. That's why, to reduce substantially the computational cost, it is possible to make the approximation to Equation 3.

On the other hand the independence of the h exists because the representation of a determined type of chromosome is completely independent from the representation of other chromosomes. The reasoning of this independence can be seen on a similar way to the chromosome independence. There does not exist a strict order of labels given past nor future history. So, thanks to both independences it is possible to approach this problem with a naive Bayes decomposition.

1.3 (2)Question 3: Briefly explain all the steps and assumptions needed to derive Eq.(9) from Eq.(7).

Equation 7 (slides) states that given a sample x, an history h' and some feedback f the optimal hypothesis $\hat{h}$ is,

$$\hat{h} = \operatorname*{argmax}_h P(h|x, h', f) \tag{4}$$

It is important to assume a deterministic feedback environment. This makes possible to define a decoding function that maps each feedback signal into its decoding $d = d(f)$. This assumption makes much easier this problem since it is not necessary to have a feedback recognition model. It is possible now to replace the feedback f with its decoding d as,

$$\hat{h} = \operatorname*{argmax}_h P(h|x, h', d) = \operatorname*{argmax}_h \frac{P(h, x, h', d)}{P(x, h', d)} \tag{5}$$

we can ignore the denominator since it does not depend on h and get the following,

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, P(h, x, h', d) = \underset{h}{\operatorname{argmax}}\, P(h')P(d|h')P(h|h', d)P(x|h) \tag{6}$$

this step can be interpreted easier taking into account the Figure 1 where a bayesian network has been used in order to model the relationships between different parameters.

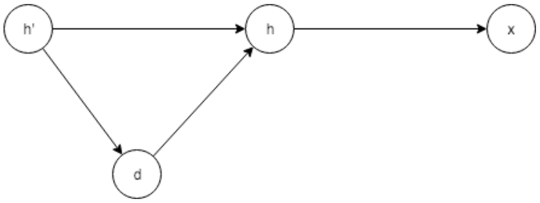

Figure 1: Bayesian network modelling relationships between variables

Finally, it is possible to clean the equation obtained on the last step, since h is obtained from both h' and d and we can consider them independent. By applying this last step we can obtain,

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, P(h')P(d|h')P(h|h', d)P(x|h) = \underset{h}{\operatorname{argmax}}\, P(x|h)P(h|h', d) \tag{7}$$

as we can observe, the result after applying all the steps explained here is the same as the equation 9 of the slides.

1.4 (3) Question 6: Briefly explain all the steps and assumptions needed to derive Eq.(19) from Eq.(7).

Similar to the previous exercise, we start from the equation 7 (slides) where feedback is taken into account. But on this question we will not assume that feedback is going to be deterministic.

As the decoding from f is not deterministic, the development of equation 7 (slides) will be slightly different form the previous exercise. The first step,

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, P(h|x, h', f) = \underset{h}{\operatorname{argmax}}\, \frac{P(h, x, h', f)}{P(x, h', f)} \tag{8}$$

It is possible to take out the denominator since it is independent to the variable we are trying to maximise. Now, we need to take into account all the possible decoding for a given f. We can add the decoding variable d marginalised as,

$$\hat{h} = \underset{h}{\operatorname{argmax}} \sum_{d} P(h, x, h', f, d) \tag{9}$$

That, taking into account the dependencies existing (see Figure 2) we can obtain,

$$\hat{h} = \underset{h}{\operatorname{argmax}} \sum_{d} P(h')P(d|h')P(f|d)P(h|h',d)P(x|h) \tag{10}$$

And simplifying it by taking out the independent probabilities and the common factor it is possible to obtain,

$$\hat{h} = \underset{h}{\operatorname{argmax}} P(x|h) \sum_{d} P(d|h')P(f|d)P(h|h',d) \tag{11}$$

This is equivalent, approximating the sum with the mode, to the equation 19 obtained in the slides. We now try to find the optimal hypothesis for a given input x and some feedback f, and the optimal decoding for a given feedback f.

$$(\hat{h}, \hat{d}) = \underset{h,d}{\operatorname{argmax}} P(d|h')P(f|d)P(x|h)P(h|h',d) \tag{12}$$

It is really interesting to observe that, the equation obtained can be seen as a multimodal fusion between a feedback recognition model and the interactive pattern recognition (IPR) model derived in the previous Question 3.

$$Feedback \rightarrow \underset{d}{\operatorname{argmax}} P(d|h')P(f|d) \tag{13}$$

$$IPR \rightarrow \underset{h}{\operatorname{argmax}} P(x|h)P(h|h',d) \tag{14}$$

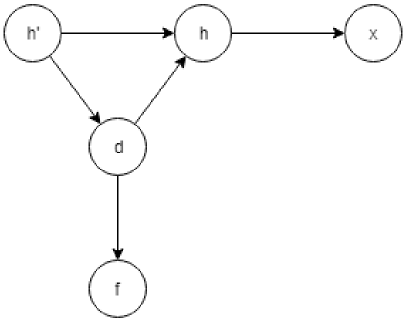

Figure 2: Bayesian network modelling relationships between variables

1.5 (3) 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 22-23 (slides) may be optimal in the case that $n =$ size of the problem. For example, in the karyotyping problem those equations would be optimal if $n = 22$ that is the number of chromosomes. That's because, on Eq. 22-23 (slides), a set of n decodings is taken into account. If there's considered the same number of decodings as the size of the problem, it is possible to have the full feedback to correct all the errors. Therefore, it is possible to calculate the optimal hypothesis.

On the other hand, equations 20-21 (slides) will never be optimal with the same conditions. Eq. 20-21 (slides) obtain first the "optimal" decoding for the feedback and, with that decoding fixed obtain the "optimal" hypothesis. The problem of this method is that there may exist a "non optimal" combination of decodings that allows to get a better hypothesis as a byproduct of both variables.

1.6 (2) Question 8: Briefly explain the concepts and main differences between Active and Passive interaction protocols.

The main difference between both active and passive interaction protocols is who decides which hypothesis element should be supervised.

In passive interactive protocol the human operator has the initiative when supervising an element from an hypothesis. This protocol can be divided in two main types regarding the order of supervision, the *left-to-right* and the *desultory*. *Left-to-right* passive protocols are those which the hypothesis elements are supervised in a fixed order. Concretely from left to the right. This type of supervision allows the system to assume that the leftmost part of the hypothesis is going to be correct and to modify the rightmost part with each user supervision. On the other hand, the *desultory* supervision is the one where the user can modify the element desired without any strict order. This supervision allows the user to choose to modify a most important error on each interaction step.

The active interactive protocol proposes a system able to take the initiative and, be the one that tells the user to supervise an element from the hypothesis. At each interaction step, the system must compute some confidence measure for each element. An approach could be that the lowest confidence element is proposed for supervision. The user, then validates whether the element is correct or needs a correction. With this modification the system computes the next predictions based on new feedback and history.

In conclusion, with passive protocols it is possible to obtain at the end a perfect output from the supervisor point of view. The main drawback of this approach is that, the human effort required is really expensive. On the other hand, with active protocols, the output may not be completely perfect from the supervisor point of view. But, since it is the system the one who takes the initiative, the human effort of this approach will be much cheaper than passive protocols.