Third Question Set

David Barbas Rebollo

Statistical Structured Prediction
February, 2021

1 Theoretical Questions

1.1 Question 1

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

Classification problems have the objective of predicting the target class of a sample out of C given classes. Usually, C is a small value. A trivial and exhaustive search is done to solve $\arg\max_{c \epsilon C} P(c|x)$.

Some examples of a classifiaction task are:

- Handwritten digit image recognition (MNIST).
- Detection of spam or non-spam e-mails.

On the other hand, structured-output prediction problems involve a possibly infinite space H. Each $h \epsilon H$ is structured into a sequence, graph, set, etc. of hypothesis elements. Therefore, the objective is to find the best structure h related to a given dataset.

Some examples of a structured-output prediction task are:

- Machine translation, translating a text from a language to another.
- Automatic summarization, generating a summirized version of a given text.

1.2 Question 2

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

The problem of karyotype recognition consists labeling 46 unsorted images of stained human chromosomes. There are 24 different labels, {1,2,...,22,X,Y}, and each label is exactly assigned to two images (except for X and Y as these are sex-dependent).

Therefore, this problem can be simplified by considering individual images instead of pairs and by using only 22 labels, excluding sex-dependent chromosomes. Meaning, there is one label per image.

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

$$P(x|h) = P(x_1, ..., x_{22}|h) = P(x_1|h)P(x_2|x_2, h)P(x_{22}|x_1, x_2, ..., x_{21}, h)$$

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

This approximation can be done because there exist and independece on both x and h. The shape of a chromosome does not depend on the shape of the others. Also, even though the subindices make the notation of the problem easier, it does not imply that chromosome need to be labeled in order. Therefore, with this approximation the computational cost is reduced substantially.

Additionally, the independence of h exists because the representation of a determined type of chromosome is completely independent from the representation of other chromosomes. This is because, there is no strict order of labels given the past or future history. So, given both independences it is possible to approach this problem with a naive Bayes decomposition.

1.3 Question 3

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

Equation 7 states that given a sample s, a history h' and a feedback f the optimal hypothesis $\hat{h}$.

$$\hat{h} = \arg\max_{h \epsilon H} P(h|x, h', f)$$

It is important to assume a deterministic feedback enviroment. This makes possible to define a decoding function that maps each feedback signal into its decoing $d = d(f)$. 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.

$$\hat{h} = \arg\max_{h \epsilon H} P(h|x, h', f) = \arg\max_{h \epsilon H} \frac{P(h, x, h', d)}{P(x, h', d)}$$

Now it is possible to ignore the denominator as it is independent from h.

$$\hat{h} = \arg\max_{h \epsilon H} P(h|x, h', f) = \arg\max_{h \epsilon H} (P(h')P(d|h')P(h|h', d)P(x|h)$$

If the equation above is represented as a bayesian network it is easier to analyze the relationships between the parameters and realize h is obtained from both h' and d. Therefore, we can consider them independent. This gives as results equation 9.

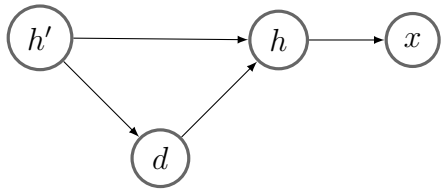

Figure 1: Bayesian network modelling the relationships between variables.

$$\hat{h} = \arg\max_{h \epsilon H} (P(h')P(d|h')P(h|h', d)P(x|h) = \arg\max_{h \epsilon H} P(x|h)P(h|h', d)$$

1.4 Question 6

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

As in the previous exercise, equation 7 is developed taking account feedback f. However, now feedback will not be assumed as deterministic.

Firstly, the denomitator of the equation below is independent to the variable being maximised. Then, it is possible to add the decoding d as a marginalised variable.

$$\hat{h} = \arg\max_{h \epsilon H} P(h|x, h', f) = \arg\max_{h \epsilon H} \frac{P(h, x, h', f)}{P(x, h', f)}$$

$$\hat{h} = \arg\max_{h \epsilon H} \sum_{d} P(h, x, h', f, d)$$

By modelling the existing dependencies as a bayesian network, it is possible to obtain the following equation.

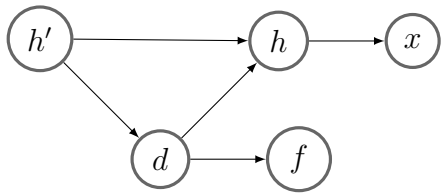

Figure 2: Bayesian network modelling the relationships between variables.

$$\hat{h} = \arg\max_{h \epsilon H} \sum_{d} P(h')P(d|h')P(f|d)P(h|h', d)P(x|h)$$

Taking out the independent probabilities and the common factor the following equation is obtained.

$$\hat{h} = \arg\max_{h \epsilon H} P(x|h) \sum_{d} P(d|h')P(f|d)P(h|h', d)$$

This is equivalent, approximating the sum with the mode, to equation 19

$$(\hat{h}, \hat{d}) = \arg\max_{h,d} P(d|h')P(f|d)P(x|h)P(h|h', d)$$

1.5 Question 7

(3) 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 karyotypingexample to illustrate your (otherwise general) responses.

The equations 22-23 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, being 22 the number of chromosomes. That is because on these equations a set of n decodings is taken into account. If number of decodings is equal to 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 21-22 will never be optimal with the same conditions. As they obtain the "first" optimal decoding for the feedback and, with that decoding the 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 Question 8

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

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

The active interactive protocol proposes a system able to take the initiative and tell 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 is to propose the supervision of the lowest confidence value element. Then, the user validates whether the element is correct or needs a correction. Given this feedback and history the system computes the next prediction.

The passive interactive protocol prososes that the user has to take initiative to supervise an element from the hypothesis. This protocol divides into 2 types of supervision. The first type is left-to-right passive protocol, which the hypothesis elements are supervised in a fixed order. With this method

allows the system to assume the leftmost part of the hypothesis is correct and the rightmost part is where the modifications occur. The second type is desultory passive protocol, it allows the user to supervise any element desired without any strict order. Meaning the user can correct the most important error on each interaction step.