Short course

A vademecum of statistical pattern recognition and machine learning

Inference and learning at a glance

Massimo Piccardi
University of Technology, Sydney, Australia

Inference and learning

- Most papers with machine learning content have a section on “inference” and one on “learning”
- *Inference* refers to estimating variables such as classes or states, given a model
- *Learning* refers to estimating the model itself from a training set
- These definitions are somewhat loose since statistical inference technically comprises both

Inference and learning

- Here, we want to learn to recognise various, common problems of inference and learning "at a glance"
- Let us assume that we are given:
  - a set of parameters, **θ**, defining the model;
  - a set of samples, **X**, from the model;
  - a set of hidden variables, **Y**, **in correspondence with X**, which could include, for instance, a set of classes
- We'll intend inference as finding the "best value" for Y and learning as finding the "best value" for θ

Preamble - 1

- Given a generic function f(x), the following properties obviously hold:

$$\arg\max_x f(x) = \arg\max_x \left(f(x) \cdot k\right)$$

$$\arg\max_x f(x) \neq \arg\max_x \left(f(x) \cdot g(x)\right)$$

- The value where the function has its maximum (argmax) is not modified by a multiplicative constant, k, but is of course modified by multiplying it by another function, g(x)

Preamble - 2

- In the case of two generic random variables, A and B, Bayes' theorem applies:

$$p(A,B) = p(A \mid B)p(B) = p(B \mid A)p(A)$$

and the properties imply:

$$\arg\max_A p(A \mid B) = \arg\max_A \left(p(A,B) = p(A \mid B)p(B)\right)$$

$$\arg\max_A p(B \mid A) \neq \arg\max_A \left(p(A,B) = p(B \mid A)p(A)\right)$$

#1

$$Y^* = \arg\max_Y p(Y \mid X, \theta)$$

- Inference: find the best values for Y, given X and θ. If Y are classes, it is also classification
- Would this problem have the same solution?

$$Y^* = \arg\max_Y p(Y, X \mid \theta)$$

- Yes. Please note that, by Bayes' theorem:

$$Y^* = \arg\max_Y \left(p(Y, X \mid \theta) = p(Y \mid X, \theta)p(X \mid \theta)\right)$$

#2

- Would this problem have the same solution?

$$Y^* = \arg\max_Y \; p(X \mid Y, \theta)$$

- No. Please note that, by Bayes' theorem:

$$Y^* = \arg\max_Y \left(p(Y, X \mid \theta) = p(X \mid Y, \theta) p(Y \mid \theta) \right)$$

This time, the two terms to maximise differ by a function, not a constant. If Y are classes, we can call the above *maximum likelihood classification*

 7

#3

$$\theta^* = \arg\max_\theta \; p(Y, X \mid \theta)$$

- Learning: maximum (joint) likelihood estimation (MLE). The hidden variables/classes are assumed known (supervised learning)
- Would this problem have the same solution?

$$\theta^* = \arg\max_\theta \; p(X \mid Y, \theta)$$

- In principle, no, because it differs by a $p(Y|\theta)$ factor. In practice, yes, since $p(Y|\theta)$ depends on a different subset of parameters than $p(X|Y,\theta)$

 8

#4

$$\theta^* = \arg\max_{\theta} \; p(Y \mid X, \theta)$$

- Learning: maximum conditional likelihood estimation (MCLE). The hidden variables/classes are again assumed known (supervised learning)

#5

$$\theta^*, Y^* = \arg\max_{\theta, Y} \; p(Y, X \mid \theta)$$

- Again, joint MLE. This time the hidden variables/classes are assumed unknown (unsupervised learning), and we find the best. Therefore, this is **joint learning and inference**. Usually, the Y* are discarded after learning and only the θ are retained for inference on future samples

#6

$$\theta^* = \arg\max_{\theta} \; p(X \mid \theta)$$

- Again, MLE. This time the hidden variables/classes are again assumed unknown (unsupervised learning), and we have marginalised them. This case is sometimes called **maximum incomplete data** (i.e. measurements only) **likelihood**. NB: the resulting θ would differ from the previous case!
- Marginalisation:

$$p(X \mid \theta) = \int_Y p(Y, X \mid \theta) dY$$

#7

$$\theta^* = \arg\max_{\theta} \; p(\theta \mid X, Y)$$

- Learning: this time the parameters are treated as a random variable! This is universally known as **maximum-a-posteriori estimation (MAPE)** (not to be confused with MAP inference!)

#8

$$\theta^*, Y^* = \arg\max_{\theta, Y} \; p(\theta, Y \mid X)$$

- MAPE, unsupervised, joint learning and inference

$$\theta^* = \arg\max_{\theta} \left(\; p(\theta \mid X) = \int_{Y} p(\theta, Y \mid X) dY \right)$$

- MAPE, unsupervised, again, learning with Y marginalised

#9

$$Y^* = \arg\max_{Y} \left(p(Y \mid X) = \int_{\theta} p(\theta, Y \mid X) d\theta \right)$$

- A more sophisticated inference, where we average over models (Bayesian treatment of the parameters)

Conclusions

- Overall, these problems differ based on:
  - if we target **θ, Y or both** in the maximisation (learning, inference or both)
  - if we assume the non-target variables to be **known or unknown** (the samples, X, are always assumed known)
  - if we assign the variables with a **probability**, or they are just conditioning values
  - if we **maximise or marginalise** the variables we do not know and are not interested in
- Proportionality constants do not affect the maximisation

