Generative models for discrete data

Lesson No. 4

João Victor da Silva Guerra

March 11, 2020

1 Introduction - notes

A generative classifier has the following form:

$$p(y = c|x,\theta) \propto \ p(x|y = c,\theta)p(y = c,\theta) \tag{1}$$

The key to using such models is specifying a suitable form for the class-conditional density $p(x|y = c,\theta)$, which defines what kind of data we expect in each class. Also, we discuss how to infer the unkown parameters θ.

2 Bayesian concept of Learning - notes

Physicological research has shown that people can learn concepts from positive examples (e.g. "this is a car", "this is a pen", etc.).

Concept learning, also known category learning, concept attainment, and concept formation, are the search for and listing of features (attributes) that helps us to classify data, and each member of a class has a set of common relevant features. If we include uncertainty about our classification, we can emulate **fuzzy set theory**, but using standard probability calculus.

Given a set of observations $\mathscr{D}$, a classic approach to induction is to suppose we have a **hypothesis space** of concepts, $\mathscr{H}$, such as: odd numbers, even numbers, all numbers between 1 and 100, etc. The subset of $\mathscr{H}$ that is consistent with the data $\mathscr{D}$ is called **version space**. As we increase out set of observations, the version space shrinks and we become increasingly certain about the concept.

2.0.1 Likelihood

Why we chose a hypothesis, when more than one is consistent with the observations? The key intuition is that we want to avoid **suspicious coincidences**.

`Example:` Given a set $\mathscr{D} = \{16,8,2,64\}$, if the true concept was even numbers, how come we only saw numbers that happened to be powers of two?

To formalize this, let us assume that examples are sampled uniformly at random from the extension of a concept (the extension of a concept is just the set of number that belong to it, e.g., the extension of $h_{even} = \{2, 4, 6, ..., 98, 100\}$). Tenenbaum (**ref** p.66 MLAPP) describes this the **strong sampling assumption**.

Given this assumption, the probability of independently sampling N items (with replacement) from h is given by:

$$p(\mathscr{D}|h) = \left[\frac{1}{size(h)}\right]^N = \left[\frac{1}{|h|}\right]^N \tag{2}$$

This crucial equation embodies what Tenenbaum calls the **size principle** (or, commonly **Occam's razor**), which means the model favors the simplest (smallest) hypothesis consistent with the data.