Maximum likelihood estimation of mixture models

Wilker Aziz

December 8, 2016

In this notes I revisit mixture models with multinomials and parameter estimation. As an application we will see a word alignment model. Along the way, I also cover some topics of the more general framework of graphical models.

1 Graphical models

Graphical models are a framework to express probability distributions over random variables (rvs).

We use rvs to capture or represent different aspects of the data such as

- observations (e.g. words in text)
- latent data: structure we believe to exist (e.g. word categories)

as well as aspects of the solution itself such as parameters of distributions.

NOTE in a frequentist treatment, parameters are seldom treated as rvs themselves. In this notes, parameters are just variables to be optimised. A Bayesian treatment of mixture models is left for future.

The framework provides us with a theory of probabilistic models which includes

- a language to represent probability distributions
- inference algorithms: probability queries
- learning algorithms: fit models to data

Graphical structure is used to encode independence statements about the problem. Nodes of the graph represent rvs and edges represent direct dependencies between rvs.

Depending on the nature of the relationship between variables we use directed or undirected graphs. Directed edges represent a causal relationship, whereas undirected edges

represent correlation. Because directed edges make statements about causality, directed graphical models are a convenient framework to express generative stories.

NOTE the terms Bayesian network and Markov network are also used to refer to directed and undirected models respectively. In terms of terminology we prefer the latter. For example, there is nothing inherently Bayesian about Bayesian networks — as a language for specifying directed graphical models they can be used in either Bayesian or frequentist settings.

1.1 Directed graphical models

Our first example of a directed graphical model is about modelling student's performance as measured by SAT scores, where we assume the student's intelligence affects his/her performance in SAT assessments. We choose to use two binary rvs: intelligence I (high vs low) and SAT score S (high vs low). Figure 1 is an example of joint distribution. Note that we could parameterise this distribution using a four-outcome multinomial distribution.

I	S	$P(I,S)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Figure 1: Example of joint distribution $P(I,S)$

Alternatively, by simple application of the chain rule, the joint probability $P(I,S)$ can be rewritten as $P(I)P(S|I)$. Now we can specify the full distribution by specifying two tables as exemplified in Figure 2: one represents the *prior* over I and the other the *conditional probability distribution* (cpd) of S given I. Now we could parameterise this distribution using 3 Bernoulli distributions (one for the prior, one for each conditional).

i^0	i^1
0.7	0.3

I	s^0	s^1
i^0	0.95	0.05
i^0	0.2	0.8

Figure 2: Example of cpds for the joint distribution $P(I,S) = P(I)P(S|I)$

In the conditional parameterisation we are explicit about the following assumption: we assume that the student's SAT score S depends directly on the student's intelligence I and we assume a causal relation. In the language of directed graphical models, we have the model depicted in Figure 3.

Figure 3: A directed graphical model for the student example.

1.1.1 Conditional independence

Let us add another variable to our problem. Suppose grade (G) which takes one of 3 values (A, B, C) is also determined by the student's intelligence alone.

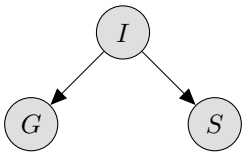

Figure 4: Example of naive Bayes model

Here we make a conditional independence statement: $P \models (S \perp G|I)$. We are assuming that intelligence is the only reason why grade and SAT scores might be correlated.

NOTE modelling assumptions are not "true" in any formal sense of the word, they are often only approximations of our true beliefs, sometimes they are motivated merely by convenience.

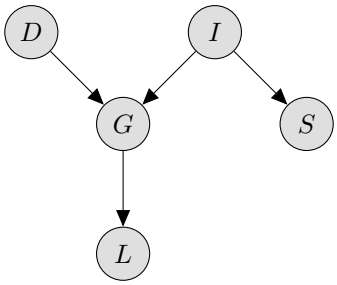

Figure 5: Student example (Koller and Friedman, 2009)

Now suppose the student is looking for an internship and needs a recommendation letter from the professor responsible for a certain course (say NLP1). We take into account the difficulty of the course D as a binary rv (difficult or easy), and we take into account

the quality of the recommendation letter L as a binary rv (good or bad). Figure 5 is an example of a directed graphical model specifying the independence assumptions we believe might be reasonable. Note 3 reasoning patterns:

- we may reason about downstream effects of a factor, this is called *causal reasoning*, e.g. $P(l^1|g^{\mathrm{C}}, l^0)$
- we may reason from effects to causes, this is called *evidential reasoning*, e.g. $P(i^1|g^{\mathrm{C}}, l^0)$
- we may reason again from effects to causes where different causes correlate with the same effect, when this happens one causal factor gives us information about the other, this is called *intercausal reasoning*, e.g. $P(i^1|g^{\mathrm{B}}, d^1)$

Note 3 local structure patterns:

- $P \models (S \perp G|I)$: upon observing intelligence, grande and SAT are independent
- $P \models (L \perp I|G)$: upon observing grade, intelligence and recommendation letter are independent
- observing G makes D and I dependent

A very strong and general results of directed graphical models is the following, for each variable X_i:

$$P \models (X_i \perp \text{NonDescendants}_{X_i} \,|\, \text{Parents}_{X_i})$$

1.1.2 Factorisation

The graphical structure encodes all necessary independence assumptions. Then, the joint distribution factorises as shown in Equation 1.

$$P(\mathbf{X}) = \prod_i P(X_i|\,\text{Parents}_{X_i}) \tag{1}$$

The individual factors $P(X_i|\,\text{Parents}_{X_i})$ are cpds or local probabilistic models. In effect, specifying a directed graphical model requires: a directed acyclic graph (DAG) and a set of local probabilistic models.

1.1.3 Parameterisation

Throughout these notes, I will model all cpds using categorical distributions (or more generally multinomials). A categorical distribution is a very natural parameterisation of a cpd and as we know for a variable defined over k possible outcomes, a categorical distribution has $k - 1$ free parameters.

Another useful result is the maximum likelihood estimate of categorical distributions for fully observable variables. Consider a dataset of observations $\mathcal{D}$ and let c and d represent an arbitrary context and decision (in the support of some rv of interest), then Equation (2) shows the MLE, where $n_{\mathcal{D}}(c, d)$ is the number of times d was caused by c in $\mathcal{D}$.

$$\theta_{c,d} = \frac{n_{\mathcal{D}}(c, d)}{\sum_{d'} n_{\mathcal{D}}(c, d')} \tag{2}$$

1.1.4 Bigram language models

Most likely you have already seen a graphical model before when you first encountered bigram language models. These are directed graphical models where each word is "caused" or generated by its preceding word.

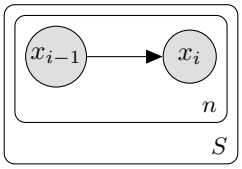

Figure 6: Bigram language model

We assume a corpus made of S sentences, each of which contains n words, and the joint distribution has a factor $P(X_i|X_{i-1})$ for each word pair in a sentence. Bigram language models are typically parameterised using categorical distributions, one distribution per conditioning context (a word type), each defined over the entire English vocabulary, thus where we denoted the size of the English vocabulary by v_E, we estimate $O(v_E \times v_E)$ parameters.

We can easily estimate these parameters by maximum likelihood given that all rvs are fully observable.

2 IBM model 1

Now we turn to a model that employs latent variables. Latent variables capture degrees of generalisation we believe to exist in the data, but are not overt.

IBM models were introduced in the 90s for word-to-word translation, they mark the birth of statistical machine translation. The first model is a simple lexical translation model, you can imagine it as learning a word-to-word dictionary, or translation table.

Let's start with the task and then propose a generative story.

We are given a sentence-aligned bilingual corpus of English and French texts and we want to generate the French data given the English data.

Imagine you were given a text

the black dog	o cao preto
the nice dog	o cao amigo
the black cat	o gato preto
the cat	o gato

now imagine the French words were replaced by placeholders

the black dog	$F_1\ F_2\ F_3$
the nice dog	$F_1\ F_2\ F_3$
the black cat	$F_1\ F_2\ F_3$
the cat	$F_1\ F_2$

but crucially, we can observe the English words as well as the French sentence length. Our task is to have a model that explains/generates the original data. To do so, we introduce a random variable called *alignment* which decorates each French position. Each alignment variable selects a position in the English sentence, and the English word occupying that position generates a French word.

This is our generative story for each sentence pair independently

1. observe an English sentence $e_1, \cdots, e_l$ and a French sentence length m
2. for each French word position j from 1 to m
   (a) select an English position a_j
   (b) conditioned on the English word e_{a_j}, generate f_j

Figure 7 is a graphical depiction of the generative story of IBM model 1. Note that IBM model 1 is a type of mixture model, where the alignment variable selects a mixture component (an English word) by selecting a position in the English sentence.

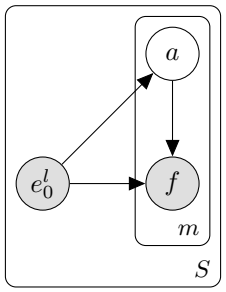

Figure 7: IBM model 1

2.1 Factorisation

For one sentence pair, from the graphical model in Figure 7, we see that the joint distribution has m independent factors of the kind $P(F_j, A_j | E = e_1^l, M = m)$.

$$P(F_1^m, A_1^m | E = e_1^l, M = m) = \prod_{j=1}^{m} P(F_j, A_j | E = e_1^l, M = m) \tag{3a}$$

$$= \prod_{j=1}^{m} P(A_j | E = e_1^l, M = m) \times P(F_j | A_j, E = e_1^l, M = m) \tag{3b}$$

2.2 Parameterisation

Another two independence assumptions made are: the alignment distribution does not depend on lexical choices, thus we can write $P(A_j | L = l, M = m)$, and the French word at position j depends on the English sentence only through the word it aligns to, thus we can write $P(F_j | E_{A_j})$. We model $P(A_j | L = l, M = m)$ uniformly in IBM model 1 and with a categorical in IBM model 2. We model $P(F_j | E_{A_j})$ with a set of categorical distributions, one per English word, each defined over the entire French vocabulary.

2.3 Inference

Likelihood:

$$P(F_1^m, A_1^m | E = e_1^l, M = m) = \prod_{j=1}^{m} P(A_j | L = l, M = m) \times P(F_j | E_{A_j}) \tag{4}$$

Marginal likelihood:
Posterior:

2.4 Parameter estimation

EM