Bayesian PLDA

Niko Brümmer

August 13, 2010

1 Introduction

The current PLDA recipe for speaker detection is a two-step process:

1. Given a development database of labelled data, make an ML point estimate of the PLDA model.
2. Given the (unlabelled) data of a detection trial, *plug in* the above point estimate to compute the posterior for the target vs non-target trial.

This note is to explore what happens if we do not use this plug-in model, but instead integrate out the parameters of the model in a fully Bayesian way.

2 Assumptions

Here we define all our assumptions about data, labels, models and priors.

2.1 Data

Let

- $\mathcal{D}_\mathrm{d}$ be all the i-vectors in the *development* data, where many speakers are present.
- Let $\mathcal{D}_\mathrm{t} = \{\ell, r\}$ be the input to the detection *trial*, where ℓ and r are i-vectors, which may be from the same speaker or from two different speakers.
- We assume throughout that the speakers of $\mathcal{D}_\mathrm{d}$ are all different from the speaker(s) of $\mathcal{D}_\mathrm{t}$.
- Let $\bar{\mathcal{D}} = \mathcal{D}_\mathrm{d} \cup \mathcal{D}_\mathrm{t}$, be the combined pool of development and trial data.
- We shall use $\mathcal{D}$ to refer in general to any of the above data sets.

2.2 Development Supervision and Trial Hypotheses

The development data is supervised w.r.t. speaker, while the detection trial is unsupervised:

- We denote the given development supervision (labelling) as θ_d. This is a partitioning of $\mathcal{D}_\text{d}$, according to speaker. No prior for θ_d is necessary, since it is given.

- Let θ_t denote the unknown speaker detection hypothesis, so that $\theta_\text{t} \in \{\mathcal{T}, \mathcal{N}\}$, where $\mathcal{T}$ is the hypothesis that ℓ and r have the same speaker; and $\mathcal{N}$ is the hypothesis that they have different speakers. Note that θ_t is also a partitioning of $\mathcal{D}_\text{t}$, where $\mathcal{T}$ is the coarsest partition and $\mathcal{N}$ is the finest partition.

- There is given a *hypothesis prior*, $\pi = (P_\mathcal{T}, P_\mathcal{N})$, such that $P_\mathcal{T} = P(\mathcal{T}|\pi) = 1 - P_\mathcal{N} = 1 - P(\mathcal{N}|\pi)$.

- Let $\bar{\theta} = \theta_\text{d} \wedge \theta_\text{t}$ denote the labelling (partitioning) of all the pooled development and trial data in $\bar{\mathcal{D}}$.

- We use θ to refer in general to any of $\bar{\theta}$, θ_d or θ_t.

2.3 Model

We assume there is a model, $\mathcal{M}$, which enables computation of any *model likelihood* of the form:

$$P(\mathcal{D}|\theta, \mathcal{M}) \tag{1}$$

Note that the ML plug-in recipe finds the plug-in model by maximizing (1) w.r.t. $\mathcal{M}$, when $\mathcal{D}_\text{d}$ and θ_d are given.

2.3.1 Conditional Independence

We assume our model is such that when its parameter, $\mathcal{M}$, is given, then we have conditional independence between speakers. In particular, since the speakers of $\mathcal{D}_\text{d}$ and $\mathcal{D}_\text{t}$ are disjoint, we have:

$$P(\bar{\mathcal{D}}|\bar{\theta}, \mathcal{M}) = P(\mathcal{D}_\text{d}|\theta_\text{d}, \mathcal{M}) P(\mathcal{D}_\text{t}|\theta_\text{t}, \mathcal{M}) \tag{2}$$

We return in section 2.4 below to a more complete representation of the dependence structure of model and data in terms of a graphical model.

2.3.2 Plug-in Recognition

If in addition to the likelihood (1), we are also given a prior for θ, say $P(\theta|\pi)$, then it is straightforward to also compute the *plug-in* posterior:

$$P(\theta|\mathcal{M}, \mathcal{D}, \pi) \tag{3}$$

It is called plug-in, because we need to plug in some fixed model $\mathcal{M}$ to do the calculation. In particular, we can form the plug-in posterior odds:

$$\begin{aligned} \frac{P(\mathcal{T}|\mathcal{D}_\text{t}, \pi, \mathcal{M})}{P(\mathcal{N}|\mathcal{D}_\text{t}, \pi, \mathcal{M})} &= \frac{P_\mathcal{T}}{P_\mathcal{N}} \frac{P(\mathcal{D}_\text{t}|\mathcal{T}, \mathcal{M})}{P(\mathcal{D}_\text{t}|\mathcal{N}, \mathcal{M})} \qquad (4) \\ &= \frac{P_\mathcal{T}}{P_\mathcal{N}} R(\mathcal{D}_\text{t}, \mathcal{M}) \qquad (5) \end{aligned}$$

where we have defined the *plug-in likelihood-ratio* $R(\mathcal{D}_\text{t}, \mathcal{M})$. This is the familiar formula: *posterior odds is the product of prior odds and likelihood-ratio.*

2.3.3 Model Posterior

To generalize the plug-in recipe to a fully Bayesian treatment, we relax the assumption that $\mathcal{M}$ is known when we process the trial and instead work with probability distributions for the model. For this we need a *model prior*, $P(\mathcal{M}|\Pi)$, as well as the means to compute the *model posterior*:

$$P(\mathcal{M}|\mathcal{D}, \theta, \Pi) \tag{6}$$

for a given dataset, $\mathcal{D}$, with given supervision, θ. Section 3 below shows how to make use of this posterior.

2.4 Graphical Model

We end this section with a summary of all of the conditional independence assumptions between data, labels, model and priors, in the form of the *graphical model* in figure 1. Here $\mathcal{D}_\text{d}$, $\mathcal{D}_\text{t}$ and θ_d are *observed variables*; θ_t and $\mathcal{M}$ are the unknown *hidden variables*; and Π and π are given fixed *priors* on the hidden variables. The first hidden variable, θ_t, is the one whose value we want to infer, while $\mathcal{M}$ is the *nuisance variable.*

In our calculations below, we shall need to determine whether some pair of variables (nodes) in the graph are conditionally independent, given some other set of variables. This is done (see e.g. section 8.22 in Bishop's book) as follows:

- Two variables, say a and b, are conditionally independent, given some set of variables, say $\mathcal{C}$, if *all paths* on the graph between a and b are *blocked.*

- A path is blocked if *any node* on the path is blocked.

- A node is blocked if either:

  - Arrows on the path meet *head-to-tail*, or *tail-to-tail* at the node and the variable at the node is in $\mathcal{C}$; or

  - Arrows on the path meet head-to-head at the node and neither the node, nor any of its descendants[1] are in $\mathcal{C}$.

In summary, if the conditions are met, then $P(a, b|\mathcal{C}) = P(a|\mathcal{C})P(b|\mathcal{C})$, or equivalently $P(a|\mathcal{C}, b) = P(a|\mathcal{C})$.

We illustrate these rules with two examples, the results of which we shall re-use in our final derivation.

2.4.1 Example

There is one path between $\mathcal{M}$ and π, which is blocked when the head-to-tail node θ_t is observed, so that:

$$P(\mathcal{M}|\theta_\mathrm{t}, \pi) = P(\mathcal{M}|\theta_\mathrm{t}) \tag{7}$$

Note that when the head-to-head node $\mathcal{D}_\mathrm{t}$, which is on the same path, is also observed, this node is *not* blocked, but the path remains blocked at θ_t. Moreover, any other variables not on the path between $\mathcal{M}$ and π do not affect their dependence. This gives for example:

$$P(\mathcal{M}|\theta_\mathrm{t}, \mathcal{D}_\mathrm{t}, \Pi, \pi) = P(\mathcal{M}|\theta_\mathrm{t}, \mathcal{D}_\mathrm{t}, \Pi) \tag{8}$$

2.4.2 Example

When $\mathcal{M}$ is given, then θ_t is independent of Π, because the path between them at $\mathcal{M}$ is head-to-tail. Also, θ_t is independent of $\mathcal{D}_\mathrm{d}$ and θ_d, because the path to them through $\mathcal{M}$ is tail-to-tail. This gives:

$$P(\theta_\mathrm{t}|\mathcal{M}, \bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) = P(\theta_\mathrm{t}|\mathcal{M}, \mathcal{D}_\mathrm{t}, \pi) \tag{9}$$

[1] A descendant of a node c is any node that can be reached from c by following the arrows.

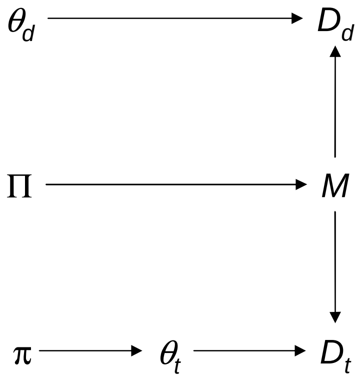

Figure 1: Graphical Model

3 Bayesian Recognition

The object of the whole exercise is to infer (compute the posterior for) θ_t, given all the data, $\bar{\mathcal{D}} = \mathcal{D}_\mathrm{d} \cup \mathcal{D}_\mathrm{t}$, the development labels θ_d and the priors Π and π. That is, we want to compute $P(\theta_\mathrm{t}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi)$.

We find the required posterior by equating the two different ways of factoring the joint posterior probability of the two hidden variables, given the observed data and parameters, $P(\theta_\mathrm{t}, \mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi)$:

$$\begin{aligned} &P(\mathcal{M}|\theta_\mathrm{t}, \bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) P(\theta_\mathrm{t}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) \\ =\, &P(\theta_\mathrm{t}|\mathcal{M}, \bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) \end{aligned} \tag{10}$$

Now we re-use the conditional independence results (8) and (9) to simplify (10), by omitting a few of the unnecessary conditioning variables:

$$\begin{aligned} &P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \theta_\mathrm{t}, \Pi) P(\theta_\mathrm{t}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) \\ =\, &P(\theta_\mathrm{t}|\mathcal{M}, \mathcal{D}_\mathrm{t}, \pi) P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi) \end{aligned} \tag{11}$$

If we summarize (11) as $AB = CD$, then A is of the form (6), which we assume to be tractable; B is the desired posterior we are solving for; C is of the (tractable) form (3); and D is an irrelevant normalization constant, because it does not depend on θ_t. To get the result in a convenient form, we recall that $\theta_\mathrm{t} \in \{\mathcal{T}, \mathcal{N}\}$ and we form the fully Bayesian posterior odds: 1

$$\begin{aligned} \frac{P(\mathcal{T}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi)}{P(\mathcal{N}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi, \pi)} &= \frac{P(\mathcal{T}|\mathcal{M}, \mathcal{D}_\mathrm{t}, \pi)}{P(\mathcal{N}|\mathcal{M}, \mathcal{D}_\mathrm{t}, \pi)} \frac{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{N}, \Pi)}{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{T}, \Pi)} \\ &= \frac{P_\mathcal{T}}{P_\mathcal{N}} R(\mathcal{D}_\mathrm{t}, \mathcal{M}) \frac{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{N}, \Pi)}{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{T}, \Pi)} \\ &= \frac{P_\mathcal{T}}{P_\mathcal{N}} R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi) \end{aligned} \tag{12}$$

where $R(\mathcal{D}_\mathrm{t}, \mathcal{M})$ is the above-defined *plug-in* likelihood-ratio and where we have newly defined the *fully Bayesian* likelihood-ratio $R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi)$. Note that the new posterior odds is again the product of prior odds and new likelihood-ratio.

3.1 Analysis

Consider the fully Bayesian likelihood ratio as defined above:

$$R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi) = R(\mathcal{D}_\mathrm{t}, \mathcal{M}) \frac{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{N}, \Pi)}{P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \mathcal{T}, \Pi)} \tag{13}$$

Firstly, notice that this formula appears strange, because $\mathcal{M}$ is in the RHS, but not in the LHS. This shows the RHS is in fact independent of $\mathcal{M}$. For practical calculation we can use any convenient value of $\mathcal{M}$.

Second, notice that the second term (the ratio) in the RHS is a correction factor applied to the plug-in likelihood-ratio. If we choose to use some plug-in model $\hat{\mathcal{M}}$, then the correction factor will only make a noticeable difference if posterior density for the model, at $\hat{\mathcal{M}}$, is noticeably different given the two alternate labellings of the trial data.

Finally, again making use of conditional independence, we can express the model posterior as:

$$P(\mathcal{M}|\bar{\mathcal{D}}, \theta_\mathrm{d}, \theta_\mathrm{t}, \Pi) = \frac{P(\mathcal{M}|\mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \Pi) P(\mathcal{D}_\mathrm{t}|\theta_\mathrm{t}, \mathcal{M})}{P(\mathcal{D}_\mathrm{t}, |\mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \theta_\mathrm{t}, \Pi)} \tag{14}$$

which allows (13) to be expressed as:

$$R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi) = \frac{P(\mathcal{D}_\mathrm{t}, |\mathcal{T}, \mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \Pi)}{P(\mathcal{D}_\mathrm{t}, |\mathcal{N}, \mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \Pi)} \tag{15}$$

or more succinctly[2] as:

$$R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi) = \frac{P(\bar{\mathcal{D}}|\mathcal{T}, \theta_\mathrm{d}, \Pi)}{P(\bar{\mathcal{D}}|\mathcal{N}, \theta_\mathrm{d}, \Pi)} \tag{16}$$

or even as:

$$R(\bar{\mathcal{D}}, \theta_\mathrm{d}, \Pi) = \frac{\int P(\mathcal{M}'|\mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \Pi) P(\mathcal{D}_\mathrm{t}|\mathcal{T}, \mathcal{M}')\, d\mathcal{M}'}{\int P(\mathcal{M}'|\mathcal{D}_\mathrm{d}, \theta_\mathrm{d}, \Pi) P(\mathcal{D}_\mathrm{t}|\mathcal{N}, \mathcal{M}')\, d\mathcal{M}'} \tag{17}$$

This confirms that the likelihood-ratio is independent of the parameter $\mathcal{M}$, which has been marginalized (integrated) out.

[2] Use $P(\mathcal{D}_\mathrm{d}|\theta_\mathrm{d}, \Pi) = P(\mathcal{D}_\mathrm{d}|\mathcal{T}, \theta_\mathrm{d}, \Pi) = P(\mathcal{D}_\mathrm{d}|\mathcal{N}, \theta_\mathrm{d}, \Pi)$.