Bayesian Scoring with Uncertainty Manipulation

Dong Wang

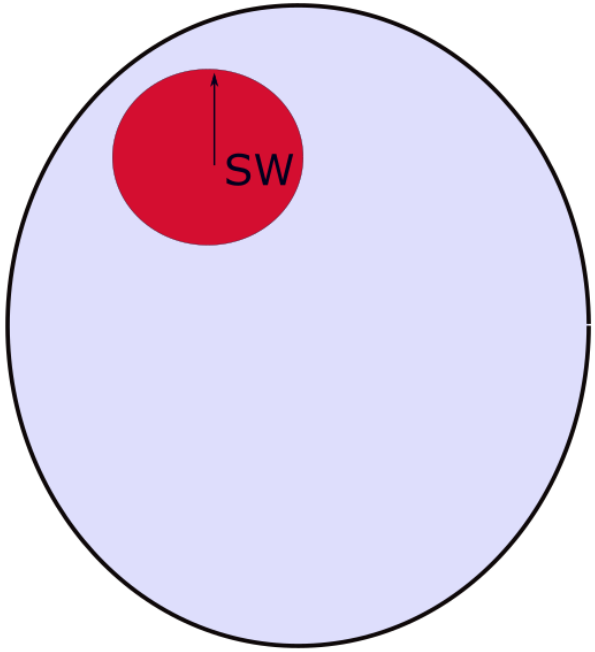
2020-03-16

MAP criterion for verification

- Verification: Given a class M, if x belongs M or not?
- Equivalently, which one is more probably?
  - A: x is generated from M
  - B: x is generated from any classes other than M?
- MAP criterion: $p(A|x)$ vs. $p(B|x)$
- $p(A|x)=p(x|A)/(p(x|A) + p(x|B))$ [with equal prior]

MAP criterion for verification

- p(x|A) = p(x|M), but what is p(x|B)?
- With a continuous prior for the class, p(x|B) need integrate all possible classes, which is equal to p(x).



Bayesian scoring

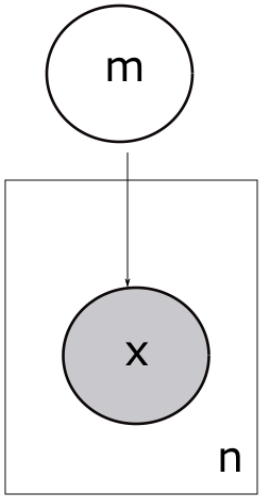
- We have x1,…,xn, now want to test if x is in the same class.
- Additional difficulty is that p(x|A) is not easy to compute, due to unknown class mean.
- A key idea is to use a distribution for the mean, rather than a value.

Two difficulties in computing p(x|A)

- How to estimate the distribution of the mean, given x1, …, xn, i.e., p(m|x1,…,xn)
- How to compute p(x|A) by giving p(m|x1,…xn)?

Mean estimation: uncertainty clamping

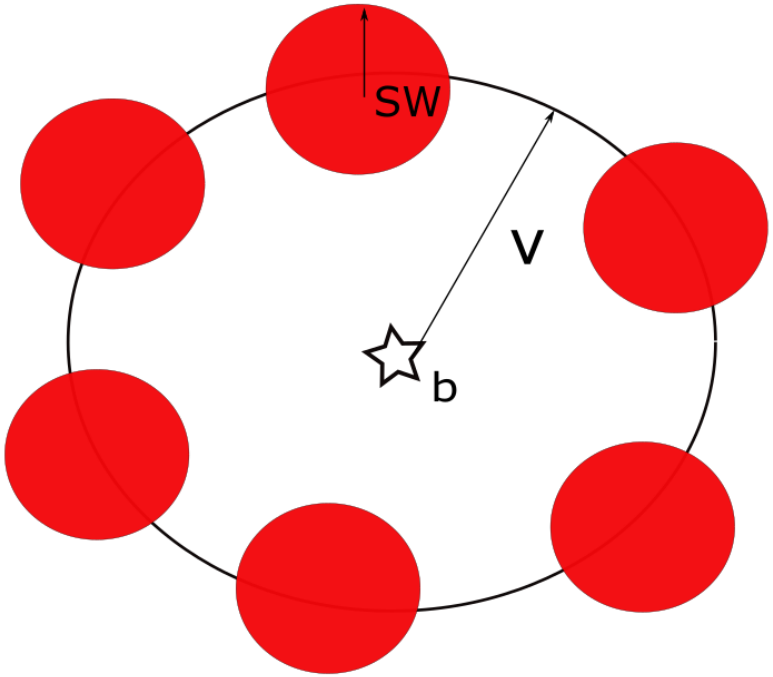
- Assume a generative model for p(m) and p(x|m), then given samples x1,…xn, it is possible to compute the posterior p(m|x1,…xn). This is called ‘inference’.
- If p(m) = N(0,SW), p(x|m)=N(m,SB)
- P(m|x1,…,xn)=N(nSB/(nSB+SW)$\bar{x}$, SB*SW/(nSB+SW))



Likelihood prediction: Uncertainty propagation

- Given p(m|x1,…xn)=N(b,v), it is easy to compute:

  p(x; x1,…xn) = N(b, SW + v)



Important message

- We separate the scoring into two phases:
  - Uncertainty clamping, that takes the observation as the evidence, to bias the generative process, to represent the specific class. (think how if no this step?)
  - Uncertainty propagation, that takes all the possibilities of the mean estimation to predict the likelihood.

Important message

- This is essentially a full Bayesian procedure [Bishop 06, Chapt 3.]

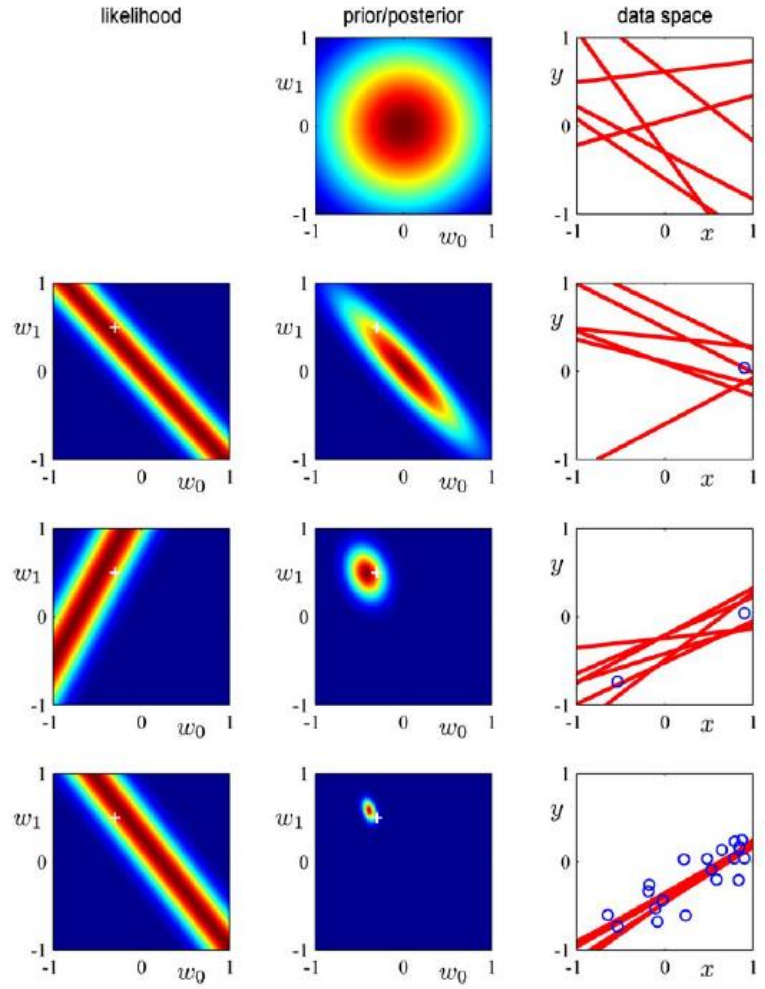


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.

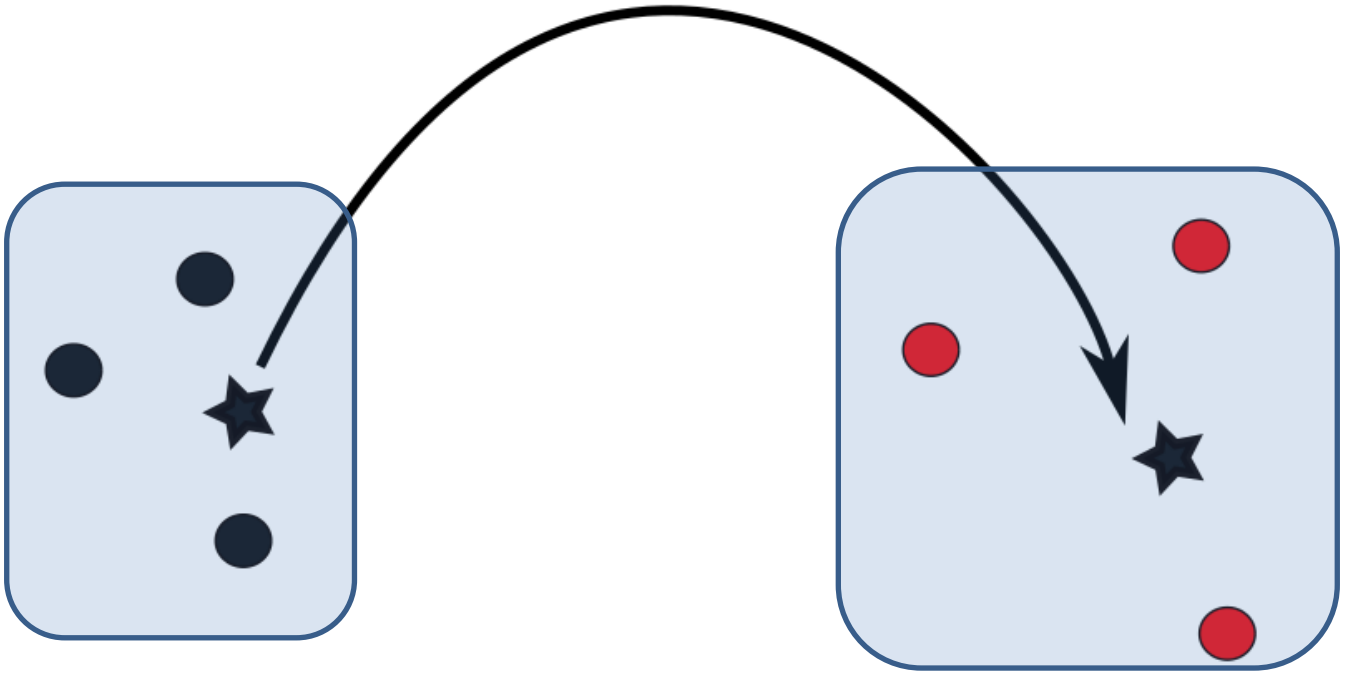
Complex uncertainty with enroll-test mismatch

- In SRE, one perform a registration, and after two months, he found he cannot be recognized.
- What happens in these two months?
  - Aging?
  - Emotion?
  - Style?
  - Environment?

Conjecture

- Statistical property mismatch between enrollment & test
  - For enroll, people often pronounce in similar ways in a single environment, usually without emotion change.
  - For test, people tend to have more complexity in terms of all the above factors.

Let’s see the picture



Observation: Distributions of enroll and test are generally different.

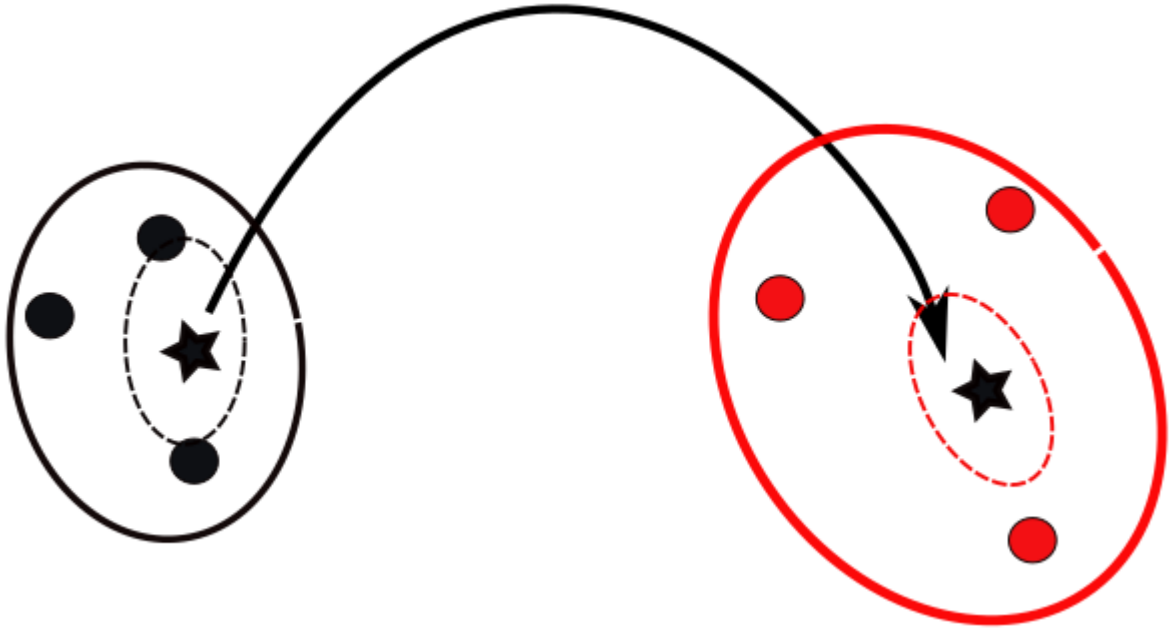
Some possible ‘solution’

- Use the enroll distribution: good for estimate mean, but bad for prediction likelihood and normalization.
- Use the test distribution: Incorrect estimate the mean, but fine for prediction likelihood and normalization.
- Map the data from enroll to test? Or vice versa? Possible, but the mapping will break the valuable statistical information.

A good solution

- We can design a transform that leads to maximum likelihood, but the likelihood is based on the statistical property of both conditions.
- We will enjoy the statistical knowledge of both conditions, and utilize transform when necessary.

Bayesian scoring for condition mismatch



- Black: enroll, red: test; dot: posterior, solid: marginal.
- Key point: the posterior will be transformed as well.

Formulation with linear transform

- Suppose a linear transform on class means:

$$\hat{\boldsymbol{x}} = \mathbf{M}\boldsymbol{x} + \boldsymbol{b},$$

- The posterior in the enroll condition:

$$p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k) = N(\boldsymbol{\mu}_k; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{I}).$$

- Map the mean, the distribution in the test condition will be:

$$p'(\hat{\boldsymbol{\mu}}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k) = N(\hat{\boldsymbol{\mu}}_k; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\mathbf{M}\bar{\boldsymbol{x}}_k+\boldsymbol{b}, \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{M}\mathbf{M}^T).$$

Formulation with linear transform

- The likelihood will be:

$$
\begin{aligned}
p'_k(\hat{\boldsymbol{x}}) &= \int p'(\hat{\boldsymbol{x}}|\hat{\boldsymbol{\mu}}_k) p'(\hat{\boldsymbol{\mu}}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k) \mathrm{d}\hat{\boldsymbol{\mu}}_\mathrm{k} \\
&= N(\hat{\boldsymbol{x}}; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\mathbf{M}\bar{\boldsymbol{x}}_k + \boldsymbol{b}, \hat{\sigma}\mathbf{I} + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{M}\mathbf{M}^T)
\end{aligned}
$$

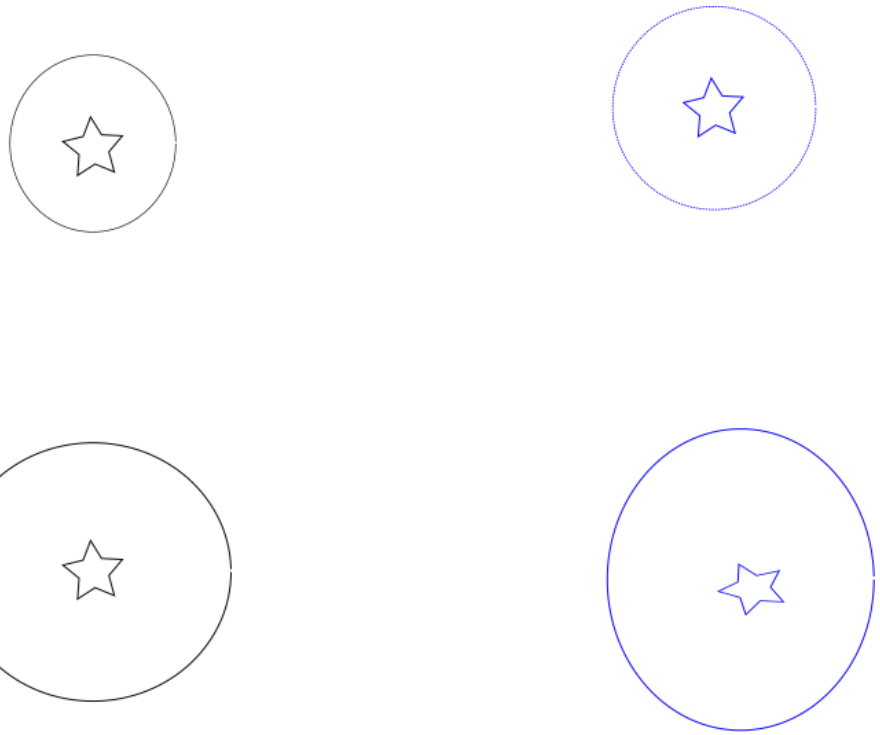
- M can be trained by maximum likelihood.

More on uncertainty manipulation

- If we only know one enrollment, but not only knows its mean, but also its distribution?
- P(x) = N(u,v)
- And we have known posterior p(m|x)=N(ax, b)
- We can derive the prediction for p(m) by marginalizing the x:
- P(m) = N(au, va^2+b)
- For a=1, we recover the uncertainty propagation base form. For v -> inf, the posterior goes to inf, indicating that if the observation is not reliable, then p(m) will be not reliable. For v=0, goes to usual posterior. Then the usual posterior is a case that the observation is extremely assured.

i-vector uncertainty for enrollment

- Considering i-vector has a distribution N(u,v), then it is possible to derive a better enrollment.
- This enrollment has usually a larger posterior, reflecting the uncertainty of the i-vector.



i-vector uncertainty for test

- How test vectors with uncertainty?
- $\iint p_1(x|m)p_1(m|x1,\dots xn)p_2(x)dmdx = \int p_1(x)p_2(x)dx$
- A correlation of two distributions.

A real case

- Suppose enhancement speech y’ for x, the estimate for clean y is p(y|y’) is Gaussian.
- The i-vector system is trained with clean speech.
- Then the author wanted to involve the uncertainty in i-vector estimation.

- Dayana Ribas and Emmanuel Vincent, AN IMPROVED UNCERTAINTY PROPAGATION METHOD FOR ROBUST I-VECTOR BASED SPEAKER RECOGNITION, ICASSP 2019.

Summary

- Since uncertainty is ubiquitous in ML, it should not be forgotten at any time.
- Especially in the deep learning era, it seems the uncertainty is largely replaced by huge data, with very weak prior.
- My conjecture is that the prior for global properties could be weak (more covered by data), but local prior would be useful.
- A possible (and interesting) ideal is a deep vector with a variance. This is essentially the DNF could do. The analysis presented here may pave the way for that kind of new embedding (Burno people call it meta embedding).

Brummer et al. Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model, 2018.