CS4786/5786: Machine Learning for Data Science, Spring 2015
3/12/2015: Lecture 14 handout: Implications of impossibility; (Gaussian) mixture models

---

Announcements

1. A2 will be posted by tomorrow, and due Friday March 27th at 5pm. It will be short, and will hence only cover basic clustering concepts.

2. Rough plan of coursework for the remainder of the term (copied from Piazza response to @127):

   > *Our plan is (still) to have the homeworks roughly correspond to one topical unit each, and the programming competitions to span more than one unit.* However, we have the interruption of spring break.
   >
   > This roughs out to approximately:
   >
   > Dimensionality reduction: A1
   >
   > Clustering: A2 (which is planned to be due Friday the 27th at 5pm, and thus be short. The TAs are bullet-testing the draft as we speak.)
   >
   > ... project ("competition") 1 .... dimensionality reduction, clustering. Not to be started during spring break (everyone deserves a break), so due not earlier than the first week after spring break.
   >
   > The remaining parts of clustering that A2 can't cover, and, more importantly, probabilistic modeling: estimate of 1-3 more assignments after A2. We don't want to guarantee that there won't be three, but we don't think that's the highest-likelihood outcome right now.
   >
   > ... project "competition") 2 .... everything in the course. According to the university registrar, *our official due date for it would be May 11th, 4:30 pm*.
   >
   > *Weighting of individual components*: still to be determined, but our basic principle is that topics of equal "depth" should have equal weight. Thus, for example, we expect the A2 that will be due before spring break to be not as heavily weighted as A1, since A2 won't cover all of clustering due to time constraints.

3. If you need a regrade request fulfilled before Wednesday in order to make an enrollment decision, please (a) send an email to your grader letting them know — we can't guarantee we'll make the deadline, but we'll try; (b) students in the College of Engineering, at least, have an option to petition for an extension to the drop deadline, but this needs to be done very soon; (c) make sure you've set your notifications on CMS for when a regrade response occurs ("one of your grades is changed").

4. As an experiment, analysis of the iClicker results from last lecture (impossibility theorem) have been posted to the course webpage. Please let us know if you find that information useful; if you do, we can generate such results for future lectures as well, but they aren't to create, and so we wouldn't proceed unless there's demand.

I. Richness property $\mathrm{Range}(f)$ = set of all partitions of X.

II. Scale-invariance property For any distance function d and any $\alpha > 0$, let $d_\alpha(t,s) \stackrel{def}{=} \alpha \cdot d(t,s)$. Then $f(d) = f(d_\alpha)$.

III. Consistency property Let d be any distance function, and let $P = f(d)$ be the partition that is output. Let d' be any distance function where:
$(t,s) \sim P$ implies $d'(t,s) \leq d(t,s)$,
and
$(t,s) \not\sim P$ implies $d'(t,s) \geq d(t,s)$.
Then, $f(d') = f(d)$.

Theorem 3.1 Let f satisfy scale-invariance and consistency. Then Range(f) is an *anti-chain* — that is, no partition in Range(f) refines any other partition in Range(f).

Theorem 3.2 For every anti-chain A, there is an f such that Range(f) $= A$ and f satisfies scale-invariance and consistency.

IV. Maximum-likelihood principle Suppose we have a generative model $P_\theta(\cdot)$ for data, where the model has parameters θ. We'll write this as $P(\cdot|\theta)$. Consider the given dataset X_{given} to be *fixed*; find the parameter setting $\hat{\theta}$ that maximizes the following function of θ: $P(X_{\text{given}}|\theta)$.

Pg. 522 of Duda, Hart and Stork (2001). Note that the notation differs from what we use in class.

10.4.1 Case 1: Unknown Mean Vectors

If the only unknown quantities are the mean vectors $\boldsymbol{\mu}_i$, then of course $\boldsymbol{\theta}_i$ consists of the components of $\boldsymbol{\mu}_i$. Equation 8 can then be used to obtain necessary conditions on the maximum-likelihood estimate for $\boldsymbol{\mu}_i$. Since the likelihood is

$$\ln p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = -\ln\left[(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}\right] - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i), \tag{14}$$

its derivative is

$$\nabla_{\boldsymbol{\mu}_i} \ln p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i). \tag{15}$$

Thus according to Eq. 8, the maximum-likelihood estimate $\hat{\boldsymbol{\mu}}_i$ must satisfy

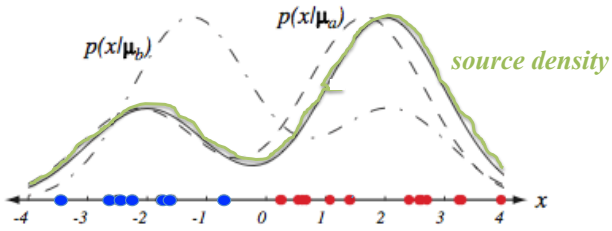
$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i) = 0, \quad \text{where } \hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c). \tag{16}$$

After multiplying by $\boldsymbol{\Sigma}_i$ and rearranging terms, we obtain the solution:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})}. \tag{17}$$

This equation is intuitively very satisfying. It shows that the maximum-likelihood estimate for $\boldsymbol{\mu}_i$ is merely a weighted average of the samples; the weight for the kth sample is an estimate of how likely it is that $\mathbf{x}_k$ belongs to the ith class. If $P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})$ happened to be 1.0 for some of the samples and 0.0 for the rest, then $\hat{\boldsymbol{\mu}}_i$ would be the mean of those samples estimated to belong to the ith class. More generally, suppose that $\hat{\boldsymbol{\mu}}_i$ is sufficiently close to the true value of $\boldsymbol{\mu}_i$ that $P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})$ is essentially the true posterior probability for ω_i. If we think of $P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})$ as the fraction of those samples having value $\mathbf{x}_k$ that come from the ith class, then we see that Eq. 17 essentially gives $\hat{\boldsymbol{\mu}}_i$ as the average of the samples coming from the ith class.

Source and estimated densities



Colors reveal the *hidden* source;
the "real data" would not be labeled like this, but look like this:

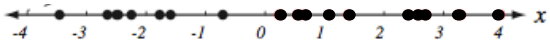
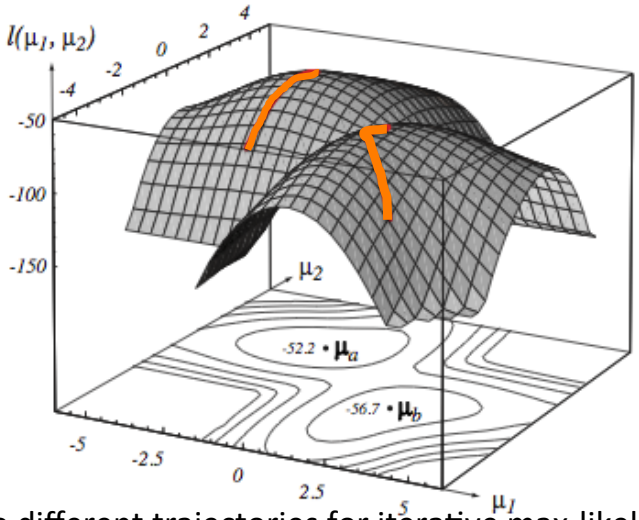


Figure adapted from Duda, Hart and Stork (2001), Fig 10.1

Log-likelihood



Orange: two different trajectories for iterative max-likelihood estimation (Starting at (0,0) would immediately converge.)

Figure adapted from Duda, Hart and Stork (2001), Fig 10.1