- Maximum likelihood framework
  - The estimation problem
  - Maximum likelihood method for the Gaussian
  - Maximum likelihood method for biased coin and die
- Mixture models
  - Motivation: piecewise-constant segmentation problem
  - Incomplete data
  - Dependence between variables
  - Maximum likelihood method with complete data

In this lecture we begin a summary of some statistical techniques that will be useful in visual grouping problems. Generally, these fall into a few groups:

1. Maximum likelihood framework ("classical" or "objective" method)
2. Bayesian framework ("subjective" method)
3. Spectral graph theory

Today we focus on the first framework. Then, motivated by a segmentation problem, we discuss mixture models and begin to apply the framework to them.

1 Maximum likelihood framework

1.1 The estimation problem

In the traditional estimation problem in statistics, we have a random variable X drawn from some population. (For example, X could represent the height of a randomly chosen Californian.) The population has some distribution, which is often assumed to have a special form. Everyone's favorite continuous distribution is the Gaussian, or normal distribution, which is expressed by[1]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

and abbreviated $N(\mu, \sigma^2)$, where μ, the mean, and σ^2, the variance, are the parameters of the distribution. Suppose we take a sample of N values of the random variable, $x_1, \ldots, x_N$

[1] What this means is that the chance that X is between x and $x + \Delta x$ approaches $f(x)\Delta x$ as Δx approaches 0.

(or more shortly (x_i)), chosen independently from one another (one says the values are "independent, identically distributed" or "i.i.d."). From the sample, we wish to estimate something about the distribution, such as μ or σ^2 in the Gaussian.

It's not clear *a priori* what is the best way to estimate a given parameter. For example, one might guess that the average of the x_i estimates the mean μ, but is this estimate better than the median? Or the average of the minimum and the maximum of the x_i?

1.2 Maximum likelihood method for the Gaussian

The maximum likelihood framework provides a way of choosing an estimator that is quite good. Specifically, it is the best estimation asymptotically as the sample size N becomes large. (Other estimators can be better for small N, but we will content ourselves with this one.) For a proof of this property, see a statistics text; here we simply show the method of determining the estimator.

For concreteness, we will assume our population is Gaussian, though the framework can be used for any distribution. If we have a sample (x_i) of size N, then the *likelihood* of observing a particular x_i is

$$P(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

The likelihood of seeing all the data is $\prod_{i=1}^{N} P(x_i)$. Viewing this as a function of the parameters, we define

$$L\left(\mu, \sigma; (x_i)\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Intuitively, this is how likely it is the population has parameters μ and σ given the sample (x_i). The *maximum likelihood estimators* for the parameters are the values $\hat{\mu}$ and $\hat{\sigma}$ that maximize $L\left(\mu, \sigma; (x_i)\right)$

It is convenient (and equivalent) to maximize the "log-likelihood" $l = \ln L\left(\mu, \sigma; (x_i)\right)$. We have

$$l = \sum_{i=1}^{N} \left[\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \right].$$

Note

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu),$$

so if $\partial l / \partial \mu = 0$, then $\mu = \sum x_i / N$. It's not hard to check this is indeed a minimum for l, so the maximum likelihood estimator for the mean of a Gaussian is $\hat{\mu} = \sum x_i / N$, the average of the sample. This is not surprising, but it does mean no other estimate (such as the median) can do better asymptotically.

With some further computation, we can see that the maximum likelihood estimator for the variance is[2] $\hat{\sigma^2} = \sum_{i=1}^{N} \sum (x_i - \hat{\mu})^2 / N$.

[2]Those acquainted with the usual definition of sample variance may wonder why N is not replaced by $N-1$ in this expression. The discrepancy arises because the maximum likelihood estimator is best only asymptotically. The estimate we get here is in fact biased, especially for small n.

1.3 Maximum likelihood method for biased coin and die

Suppose we flip a coin that turns up heads with probability θ. We'll now find the maximum likelihood estimator for θ. Take a sample of N flips (x_i). Say we get n_1 heads. Then the likelihood is $L(\theta;(x_i)) = \theta^{n_1}(1-\theta)^{N-n_1}$ and the log-likelihood is $l = n_1 \ln\theta + (N-n_1)\ln(1-\theta)$. One can check that l is minimized when

$$\frac{\partial l}{\partial \theta} = \frac{n_1}{\theta} - \frac{N-n_1}{1-\theta} = 0,$$

or $\theta = n_1/N$. Again, this is an intuitively appealing answer.

Similarly, we can estimate the probabilities $\pi_1, \ldots, \pi_6$ of a biased die rolling numbers $1, \ldots, 6$. We observe n_1 ones, n_2 twos, etc. If we do a similar computation (remembering to maximize likelihood subject to the constraint $\sum \pi_i = 1$), we get $\hat{\pi}_k = n_k / \sum n_k$.

2 Mixture models

2.1 Motivation: piecewise-constant segmentation problem

Suppose we have an image of pixels, each characterized by a single number indicating its brightness. (For simplicity, we ignore color at the moment.) The image consists of various regions representing objects that are *piecewise constant* in brightness. However, the image will have "noise," so each region will have some fluctuations. We wish to identify these regions. This is the piecewise-constant segmentation problem.

2.2 Incomplete data

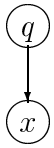
We assume that in region k of the image, the pixel brightness is $N(\mu_k, \sigma_k^2)$. Now if we already knew the exact position of the regions, we could use the method we've just described to estimate brightnesses in each region. Of course, in the segmentation problem, we don't have this information. We can think of each pixel as having two pieces of information: its brightness x_i and a label q_i saying to which region it belongs. (For now, we will assume we know there are K regions.) We know the first piece of information, but not the second. That is, we are solving an estimation problem with incomplete data.

As another example, we could take a set of observations (x_i, q_i) for N people, where each x_i and q_i represent the height and sex of a person, respectively. If we had only height measurements, how could we determine the sex of the people measured? One process for solving such problems is called *expectation maximization* (EM), which will be covered in the next lecture.

2.3 Dependence between variables

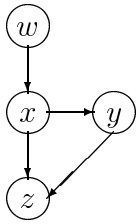
In the height/sex example, we have a joint probability distribution $P(x, q)$ with two variables. We may wish to model this distribution by considering the distribution among males and among females separately. That is, we write $P(x, q) = P(q)P(x|q)$ (where $P(q)$ is the

distribution of q alone and $P(x|q)$ is the distribution of x given the specific choice of q) and represent this graphically:



More generally, we can have a distribution $P(x, q)$ of two variables with q taking values $1, 2, \ldots, K$ with probability $\pi_1, \ldots, \pi_K$, and $P(x|q = k) = f_k(x; \phi_k)$, where f_k is the distribution of x given $q = k$, and ϕ_k are its parameters. This is called a *mixture model*.

Incidentally, note that the graphical representation above can be extended to more variables. Thus



indicates that we can write $P(w, x, y, z) = P(w)P(x|w)P(y|x)P(z|x, y)$.

2.4 Maximum likelihood method with complete data

Say we wish to estimate the parameters π_k and ϕ_k in a mixture model given a sample of observations (x_i, q_i). As we have seen, our interest is in a sample with incomplete data (missing q_i in the segmentation problem), but the complete data situation is still useful to develop notation and familiarity with the problem.

We'll find it convenient to use a (perhaps strange) notational convention: regard the value of q as a vector with K components, and represent the value k by putting a 1 in the kth component and 0's in all other components. Thus the column vector $(0, 0, 1, 0 \ldots, 0)^t$ represents 3, and q^k, the kth component of q, is 0 unless q represents k.

Now the distribution function may be written

$$P(x, q; \pi_k, \phi_k) = \prod_{k=1}^{K} [\pi_k f_k(x; \phi_k)]^{q^k}$$

so the log-likelihood function is

$$\begin{aligned} l &= \ln L(\pi_k, \phi_k; x_i, q_i) \\ &= \ln \prod_{i=1}^{N} \prod_{k=1}^{K} [\pi_k f_k(x_i; \phi_k)]^{q_i^k} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} q_i^k \ln \pi_k + (\text{terms not involving } \pi_k) \\ &= \sum_{k=1}^{K} n_k \ln \pi_k + (\text{terms not involving } \pi_k) \end{aligned}$$

where n_k is the number of instances of q_i's taking value k. Maximizing this with respect to the π_k and subject to the restraint $\sum \pi_k = 1$ we see the π_k must be chosen in the same proportions as the n_k. (For example, substitute $\pi_2 = 1 - \pi_1 - \pi_3 - \cdots - \pi_K$ into the last expression, and note $\partial l/\partial \pi_1 = 0$ implies $n_1/\pi_1 - n_2/\pi_2 = 0$. This can be done for all pairs of π_i.) Hence the maximum likelihood estimators are $\hat{\pi}_k = n_k / \sum n_k$, which is what we would expect. One could perform a similar procedure to maximize the parameters ϕ_k (given a choice of distributions f_k).