ANLY-601
Advanced Pattern Recognition

Spring 2018

L10 – Parameter Estimation

GEORGETOWN UNIVERSITY

Parameter Estimation

Maximum Likelihood and Bayes Estimates

The following lectures expand on our earlier discussion of parameter estimates, introducing some formal grounding. (A good supplemental source for this is chapter 2 of *Neural Networks for Pattern Recognition*, Chris Bishop.)

We'll discuss parametric density estimates in more detail, including mixture models for densities.

Parametric Density Models

A model of specific functional form.

A small number of parameters that are estimated from data.

e.g. Normal distribution

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{1}{2\sigma^2} (x - \mu)^2$$

Data - D = {x1, x2, x3, ..., xm}

Parameter estimates

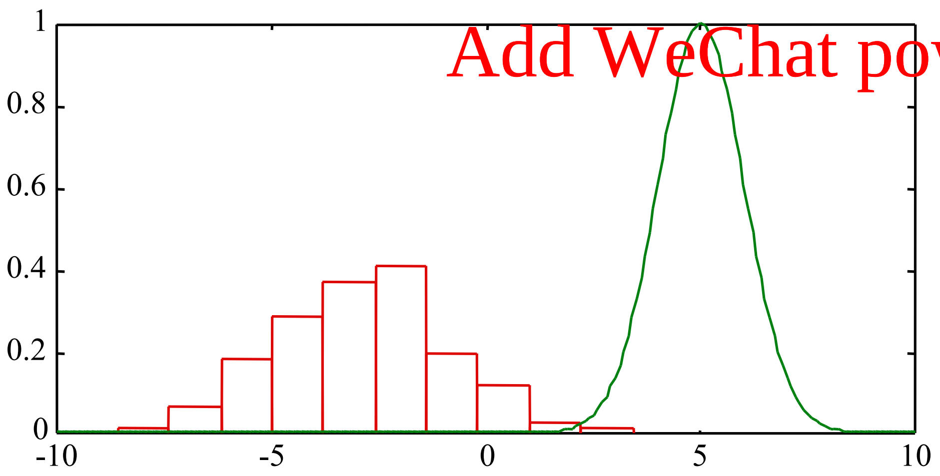
$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i \quad , \quad \hat{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(x_i - \hat{\mu}\right)^2$$

but where did these forms for the estimators come from?

Maximum Likelihood Estimation

Question -- what's the probability that the dataset D occurs, given the form of the model density?
We assume each of the x_i are sampled independently from the underlying (normal in this example) distribution, then

$$p(D \mid \mu,\ \sigma^2) = p\left(\{x_1,\ \ldots\ x_m\} \mid \mu,\ \sigma^2\right) = \prod_{i=1}^{m} p(x_i \mid \mu,\ \sigma^2)$$



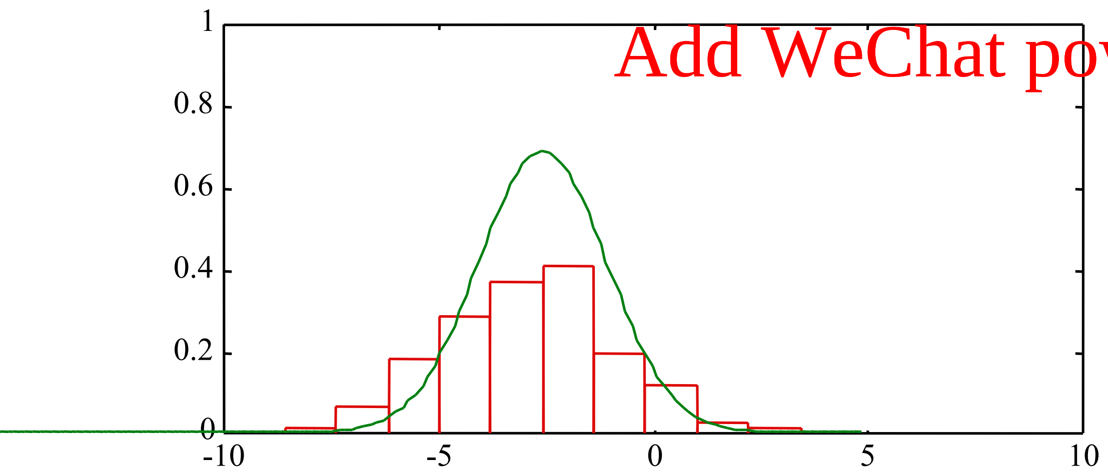
If the data (histogram) and model look like this, the data will have **low probability** given the model (data at the tails of the model distribution).

Maximum Likelihood Estimation

Question -- what's the probability that the dataset D occurs, given the form of the model density?

We assume each of the x_i are sampled independently from the underlying (normal in this example) distribution, then

$$p(D \mid \mu,\ \sigma^2) = p\left(\{x_1,\ \ldots\ x_m\} \mid \mu,\ \sigma^2\right) = \prod_{i=1}^{m} p(x_i \mid \mu,\ \sigma^2)$$

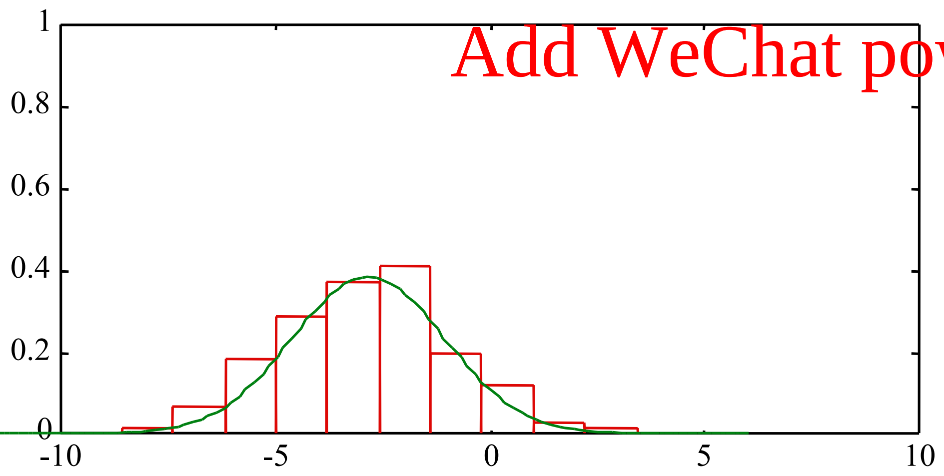


We can adjust the model mean so that the data has higher probability under the model. But this model still has low tails where there's plenty of data, because the peak is too sharp.

Maximum Likelihood Estimation

Question -- what's the probability that the dataset D occurs, given the form of the model density?
We assume each of the x_i are sampled independently from the underlying (normal in this example) distribution, then

$$p(D \mid \mu, \sigma^2) = p\left(\{x_1, \ldots x_m\} \mid \mu, \sigma^2\right) = \prod_{i=1}^{m} p(x_i \mid \mu, \sigma^2)$$



Here's a Gaussian model that chooses the mean and variance so as to <u>maximize the data likelihood under the model</u>.

Maximum Likelihood Estimation

So, we adjust the model parameters to maximize the data likelihood. Since the log is monotonic in its arguments, and we often deal with model distributions from the exponential family, it's convenient to maximize the log-likelihood.

$$L = \ln\ p(D \mid \mu, \sigma^2) = \sum_{i=1}^{m} \ln\ p(x_i \mid \mu, \sigma^2) = \sum_{i=1}^{m} \left[-\tfrac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\,\sigma^2} (x_i - \mu)^2 \right]$$

$$\left.\frac{\partial L}{\partial \mu}\right| = 0 = \tfrac{1}{\sigma^2} \sum_{i=1}^{m} (x_i - \mu) \quad \Rightarrow \quad \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

$$\left.\frac{\partial L}{\partial \sigma^2}\right| = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \tfrac{1}{m} \sum_{i=1}^{m} (x_i - \hat{\mu})^2$$

Note σ^2 is biased.

Data Distributions and Cost Functions

Regression - Minimizing mean square error between the data and a regression curve is equivalent to maximizing the data likelihood under the assumption that the fitting error is Gaussian.

The data is the sequence of (x,y) coordinates. The data y values are assumed Gaussian distributed with mean g(x). That is

$$y = g(x) + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

The data likelihood is

$$p(\{y_i\} \mid \{x_i\}; g(x), \sigma^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp -\left[\frac{1}{2\sigma^2} \left(y_i - g(x_i) \right)^2 \right]$$

Data Distributions and Cost Functions
Regression

Maximizing the data log-likelihood *L* with respect to *g(x)*

$$L = \log\left(p\left(\{y\} \mid \{x\}; g(x), \sigma^2\right)\right) = \sum_{i=1}^{m} \frac{-1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - g(x_i)\right)^2$$

is equivalent to minimizing the sum-squared fitting error with respect to *g(x)*.

$$E = \sum_{i=1}^{m} \frac{1}{2\sigma^2}\left(y_i - g(x_i)\right)^2$$

Data Distributions and Cost Functions
Classification

For a (two-class) classification problem, it's natural to write the data likelihood as a product of Bernouli distributions (since the target values are *y = 0 or 1* for each example)

$$L = p\left(\{y_i\} \middle| \{x_i\}; \alpha(x)\right) = \prod_{i=1}^{m} \alpha(x_i)^{y_i} \left(1 - \alpha(x_i)\right)^{(1-y_i)}$$

where $\alpha(x)$ is the probability that for the feature vector x, the class label is 1 (rather than 0).

Maximizing this data likelihood is equivalent to minimizing its –log, the cross entropy error

$$E = \sum_{i=1}^{m} y_i \log\left(\alpha(x_i)\right) + (1 - y_i) \log\left(1 - \alpha(x_i)\right)$$

Bayesian Estimation and Parameter Posterior Distributions

Maximum likelihood estimation --- there exits an actual value of the parameters Θ_0, that we *estimate* by maximizing the probability of the data conditioned on the parameters

$$\Theta_0 = \arg\max_{\Theta} p(D \mid \Theta)$$

An arguably more natural approach is to find the most probable values of the parameters, conditioned on the available data. Thus, parameters are regarded as random variables with their own distribution -- the posterior distribution

$$p(\Theta \mid D) = \frac{p(D \mid \Theta)\ P(\Theta)}{p(D)}$$

where $P(\Theta)$ is the *prior* on Θ

Maximum A Posterior Estimation

Maximizing the log of the posterior, with respect to the parameters, gives the maximum a posterior (or MAP) estimate

$$\hat{\Theta} = \arg\max_{\Theta} \left[\log\left(p(D \mid \Theta)\right) + \log\left(P(\Theta)\right) \right]$$

The prior distribution $P(\Theta)$ is <u>chosen</u> . Notice that if the prior is independent of Θ (flat prior), the MAP and maximum likelihood estimates are the same.

Convenient to choose the prior distribution $P(\Theta)$ so that the consequent posterior $p(\Theta \mid D)$ has the same functional form as $P(\Theta)$. (The proper form depends, of course, on the form of the likelihood function $p(D \mid \Theta)$.)

Called <u>conjugate priors</u>.

Example: Posterior Distribution of the Gaussian Model's Mean

Suppose the data is Gaussian, and the variance is known, but we just want to estimate the mean. The conjugate prior for this is a Gaussian. The posterior on the mean is

$$p\left(\mu \mid D, \sigma^2\right) = \frac{p\left(D \mid \mu, \sigma^2\right) p(\mu)}{p(D)}$$

$$= \frac{1}{p(D)\left(\sqrt{2\pi\sigma^2}\right)^m \sqrt{2\pi\lambda^2}} \exp - \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2 + \frac{1}{2\lambda^2} (\mu - \mu_0)^2 \right\}$$

where, λ and μ_0 are the variance and mean of the prior distribution on μ.

Example: Posterior Distribution of the Gaussian Model's Mean

After some algebraic manipulation, we can rewrite the posterior dist. as:

$$p\left(\mu \mid D, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{1}{2\sigma_\mu^2}\left(\mu - \bar{\mu}\right)^2\right]$$

with the posterior mean and variance given by (show this!)

$$\bar{\mu} = \frac{m\lambda^2}{m\lambda^2 + \sigma^2} \frac{1}{m}\sum_{i=1}^{m} x_i + \frac{\sigma^2}{m\lambda^2 + \sigma^2} \mu_0$$

$$\sigma_\mu^2 = \frac{\sigma^2 \lambda^2}{m\lambda^2 + \sigma^2}$$

GEORGETOWN UNIVERSITY

Example: Posterior Distribution of the Gaussian Model's Mean

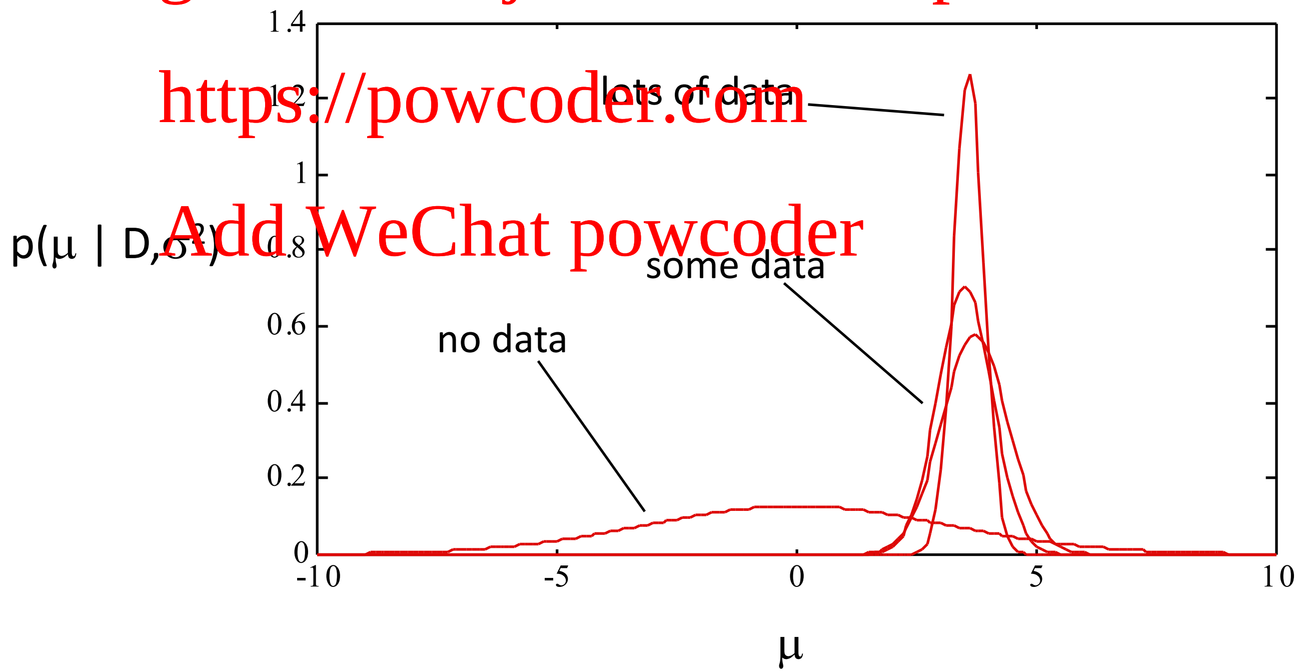
$$\bar{\mu} = \frac{m\lambda^2}{m\lambda^2 + \sigma^2} \frac{1}{m} \sum_{i=1}^{m} x_i + \frac{\sigma^2}{m\lambda^2 + \sigma^2} \mu_0$$

$$\sigma_\mu^2 = \frac{\sigma^2 \lambda^2}{m\lambda^2 + \sigma^2}$$

Note that for $m >> \sigma^2 / \lambda^2$ the posterior mean approaches the sample mean (the ML estimate), and the posterior variance becomes small.

Example: Posterior Distribution of the Gaussian Model's Mean

Without data, *m=0*, the posterior is just the original prior on μ . As we add samples, the posterior <u>remains Gaussian</u> (that's the point of a conjugate prior) but it's mean and variance change in response to the data.



1789

GEORGETOWN UNIVERSITY