Parameter Estimation: Maximum Likelihood Estimation and Bayesian Learning

Prof. Richard Zanibbi

Maximum Likelihood Estimation

Assume

Likelihood density for each class has known form, given by a parameter vector theta, e.g.

$$p(\mathbf{x}|\omega_j) \;\sim\; N(\mu_j, \Sigma_j) \qquad\qquad \theta \text{ contains } \mu_j, \Sigma_j$$

$$p(\mathbf{x}|\omega_j, \theta_j)$$

Task

Estimate theta from training samples

Definition of MLE

Likelihood of theta w.r.t. a sample set

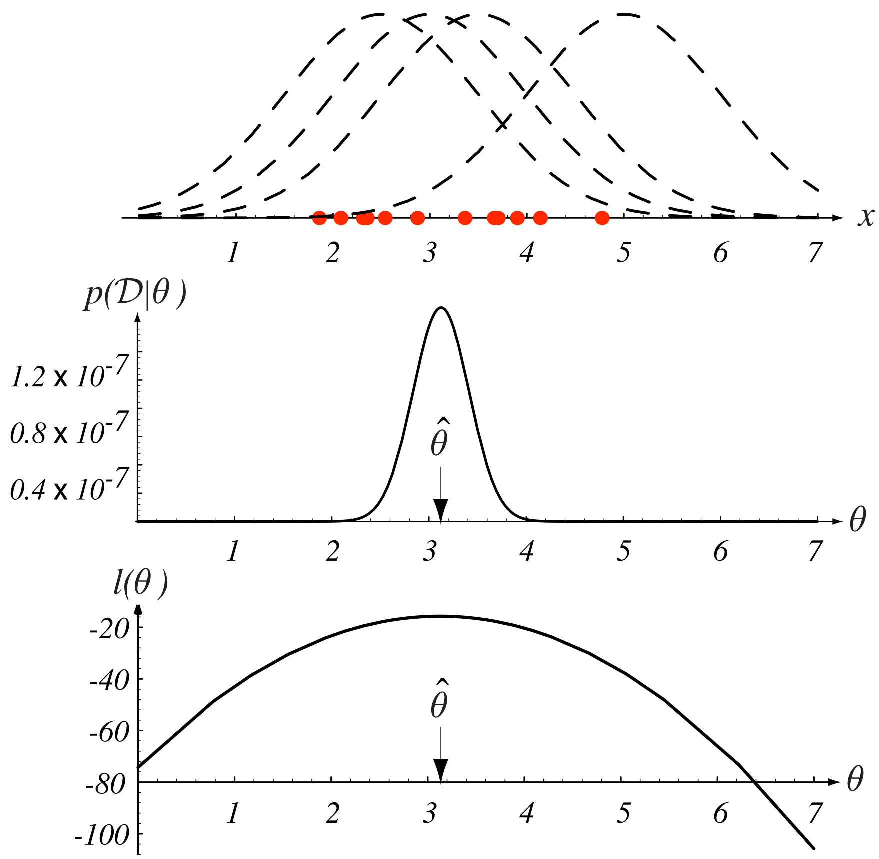
Assuming samples are independent and identically distributed:

$$p(D|\theta) = \prod_{k=1}^{n} p(\mathbf{x_k}|\theta)$$

Maximum-Likelihood Estimate of theta

The vector which maximizes p(D|theta); "best agrees" with the observed samples

Example: Maximum Likelihood Estimate of the Mean



Finding the MLE

Log-likelihood $$l(\theta) = \sum_{k=1}^{n} \ln p(\mathbf{x_k}|\theta)$$

Gradient of Log-Likelihood

(Assuming p(D|theta) differentiable, well-behaved!)

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(\mathbf{x_k}|\theta)$$

Solve for MLE of theta using: $$\nabla_\theta l = 0$$

May have multiple solutions; risk of local minima or inflection points

R·I·T

MLE Estimate of the Mean

Assuming multivariate normal, MLE for the mean must satisfy:

$$\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x_k} - \hat{\mu}) = 0$$

Multiply and rearrange to obtain (drum roll please):

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x_k}$$

R·I·T

MLE Estimate for Mean and Covariance

$$\theta_1 = \mu, \theta_2 = \sigma^2 \text{ (univariate) } \theta_2 = \Sigma \text{ (multivariate)}$$

Conditions:

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

Substitute estimates for thetas, rearrange:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x_k}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

R·I·T

Bias

Our variance, covariance estimates are biased

i.e. expected value over all data sets of size n is *not* the estimated value.

Fix (simple)

Average over n-1, not n for estimated value

Unbiased Estimators

Absolutely Unbiased

Estimator is unbiased for *all* distributions

Asymptotically Unbiased

Estimator tends towards becoming unbiased as n (# sample) becomes large

- Often acceptable for PR problems with large training data available

Effect of Invalid Model (assumed distribution)

Will the theta obtained by MLE produce the best classifier over the assumed space of models?

No.

- If model selection is poor, cannot be certain that inferred classifier is the best possible in our model set (space)

Example: Bayesian Learning of the Mean

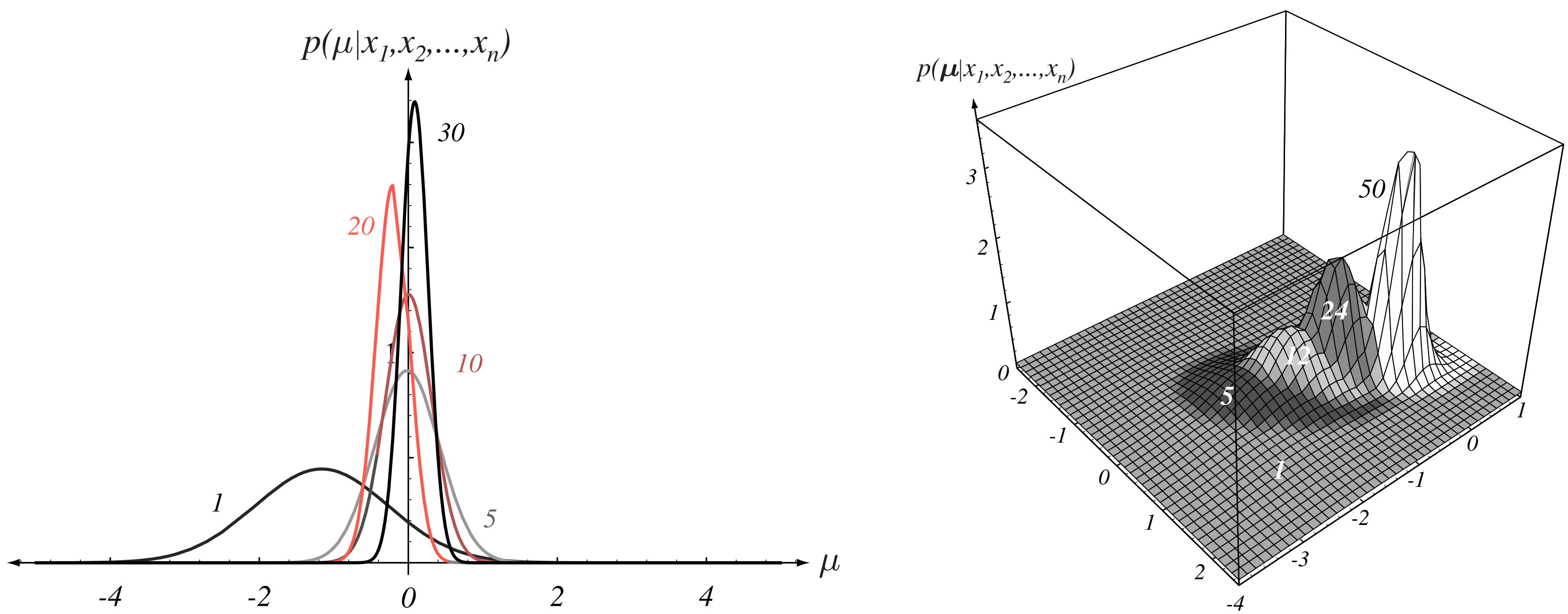


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

R·I·T

Adding Features to Better Separate Classes

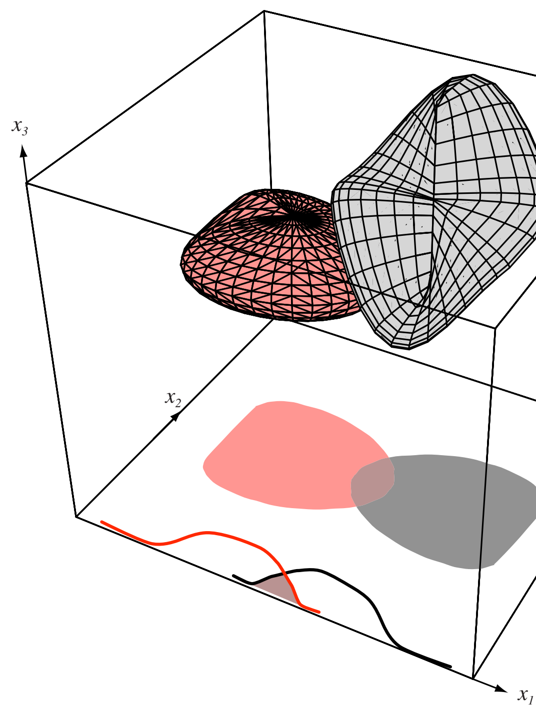


FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Overfitting: An Example

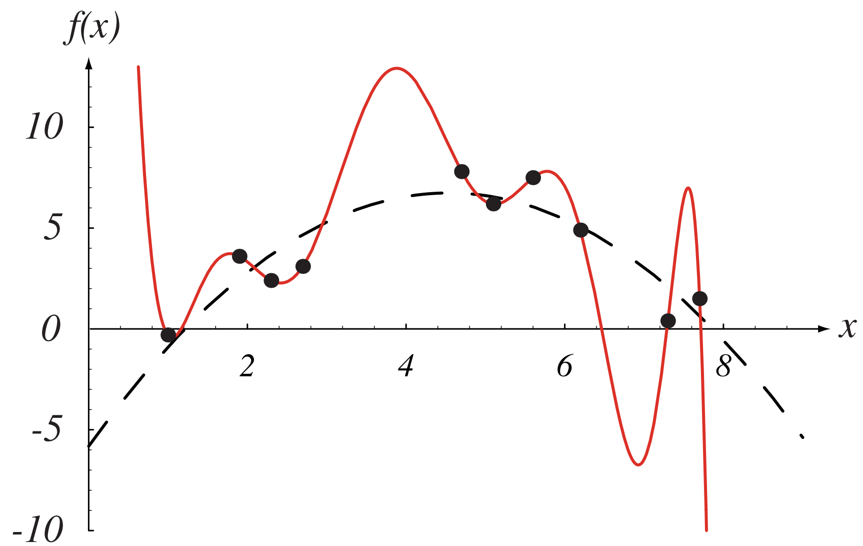


FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because it would lead to better predictions for new samples. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

R·I·T