Parameter Estimation I: Maximum Likelihood

Matthew Stone
CS 520, Spring 2000
Lecture 4

Parameter Estimation

- **Fitting continuous values from data**
- **Two kinds of problems**
  - Extensions of classification where you estimate continuous instantaneous state
    - E.g., in computer vision
  - Probabilistic learning where you describe the distribution of examples

Parameter Estimation as Learning: Motivation

- **We know how to build optimal classifiers, given exact probabilistic models**
- **But these are more realistic givens:**
  - Vague general knowledge about problem
  - Design samples or training data representative of the patterns to classify
- **How do we use this to design a classifier?**

Parameter Estimation as Learning

- **Use training data to estimate unknown probabilities and probability density functions**

Easy Case: Probabilities

- **Provided you have enough data:**
  - N samples of which c_i are in ω_i

    estimate $P(\omega_i)$ as c_i / N

Hard Case: Density Functions

- **Data is always sparse**
  - Continuous events never happen twice
  - Curse of dimensionality
    - With more features you need exponentially more samples to cover events equally closely with training data

Solution: Modeling

- **Constrain the density estimation by making assumptions about the form of the distribution**
- **Rather than learning some function, estimate the parameters of the modeled distribution**

Example

- **Normal distribution**
  - Use general knowledge that measurement involves lots of independent noise
  - Assume

$$p(x \mid \omega_i) \sim N(\mu, \sigma^2)$$

  - To describe density, just need to estimate μ and σ – not the whole function

Two Procedures – I

- **Maximum likelihood**
  - The distribution parameters are fixed
  - Know no expectations about likely values
  - Base estimate on the assumption that the training data is representative
- **Find values of parameters that make the training data as likely as possible**

Two Procedures – II

- **Bayesian estimation**
  - We have expectations about parameters
    - Expectations are expressed as prior density on parameter values
    - We want to combine these expectations with measurements (training data)
- **Use Bayes's formula to derive a posterior**
  - First for parameter values
  - Then for future measurements

Background

- **Describing training data mathematically**
  - We have *c* sets of samples

    Each $\mathcal{D}_i$ contains $\mathbf{x}_1 \cdots \mathbf{x}_{n_i}$ iid by $p(\mathbf{x} \mid \omega_i)$
  - iid: independent, identically distributed

Background

- **Describing training data mathematically**
  - Density takes a known parametric form determined by parameter vector θ

    E.g. for $p(\mathbf{x} \mid \omega_i) \sim N(\mu_i, \Sigma_i)$

    θ gives components of μ_i, Σ_i
- **Use info from samples to provide estimates of parameters**

Background

- **Simplifying assumptions and notation**

  $\mathcal{D}_i$ gives no info about θ_j unless $i = j$

➥ ***c* separate problems**

  – Use n values iid by $p(\mathbf{x}|\theta)$ to estimate θ

Maximum Likelihood

- **Likelihood**

$$p(\mathcal{D} \mid \theta) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \theta)$$

- **Maximum likelihood estimate**
  - The value of θ that maximizes $p(\mathbf{x}|\theta)$
  - The value of θ that in some sense best supports the data

Parenthesis

- **We'll adopt the usual trick of using log likelihood to facilitate reasoning about exponential (e.g., normal) distributions**

$$l(\theta) = \ln p(\mathcal{D} \mid \theta)$$

Deriving MLEs

- **Want**

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

- **"Try differentiating"**

$$\text{solve } \nabla_{\theta} l = 0$$

$$l(\theta) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \theta)$$

$$\nabla_{\theta} l = \sum_{k=1}^{n} \nabla_{\theta} \ln p(\mathbf{x}_k \mid \theta)$$

An Example

- **Normal distribution, one feature:**

$$p(x_k \mid \theta) \sim N(\theta_1, \theta_2)$$

$$\ln p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

An Example

- **To maximize, solve:**

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

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

An Example

- **We derived solution as:**

$$\hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$$

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