ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`i at Manoa

Kyungim Baek

Announcement

- Homework assignment #1 **due today by 5:00 PM**
- Project proposal
  - Due: **Wednesday September 23, by 5:00 PM**
  - Presentation (~ 10 minutes)
    - Monday (9/28):
      - BJ, Thomas, Tyson, Sharif, Danny, Jeremy
    - Wednesday (9/30):
      - Tetsuya, Kelly, Nurit
    - See the evaluation criteria for presentations posted in the course website
- **Exam I: Wednesday, October 7**

Proposal Presentation

- **Title:** A short descriptive title for the project
- **Project overview:**
  - Description of the problem and main idea of the project
  - Motivation and relevant previous work
  - What will come out of this project? (project goals)
- **Technical and experimental plan:**
  - Data: Going to use existing dataset? Will collect a new dataset? Where is the source of the dataset or how will you collect a new dataset?
  - Algorithms to be explored
  - How are you going to incorporate existing techniques? If you are proposing a new approach, what is the basic idea?
  - How is the project going to be evaluated?
- **A schedule of work:** Rough milestones
- **Bibliography:** Key references

Lecture 7

- Parametric Density Estimation
  - Bayesian Parameter Estimation
    - Formal discussion in general terms
    - Univariate Gaussian Case
    - Multivariate Gaussian Case
    - Recursive Bayes

Bayesian Estimation (I)

- Bayesian learning to pattern classification problems
  - In MLE θ was supposed to be fixed
  - In Bayesian estimation θ is a random variable
  - The computation of posterior probabilities $P(\omega_i|\mathbf{x})$ lies at the heart of Bayesian classification
  - **Goal:** compute $P(\omega_i|\mathbf{x}, D)$
    - Given the sample D, Bayes formula can be written

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j \mid D)}$$

Bayesian Estimation (II)

- This is a supervised problem so far:
  - $P(\omega_i) = P(\omega_i|D)$: assume that the prior probability is known or obtainable from the training samples
  - $D = \{D_1, D_2, \ldots, D_c\}$

$$p(\mathbf{x} \mid \omega_i, D) = p(\mathbf{x} \mid \omega_i, \{D_j\}_{j=1,\ldots c}) = p(\mathbf{x} \mid \omega_i, D_i, \{D_j\}_{j \neq i})$$
$$= p(\mathbf{x} \mid \omega_i, D_i)$$

  - Then,

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D_j)P(\omega_j)}$$

Bayesian Estimation (III)

- Notationally:

$$p(\mathbf{x}|\, \omega_i, D_i) \rightarrow p(\mathbf{x}|D)$$

- Now our problem is to <u>estimate density for **x** given the data *D*</u>
- We assume the form of $p(\mathbf{x})$ – the source density for D:

$$p(\mathbf{x}) \rightarrow p(\mathbf{x}|\theta)$$

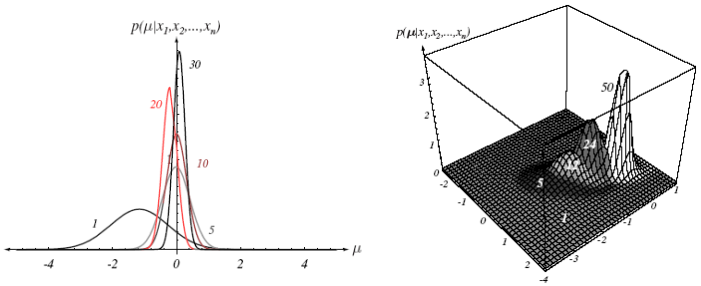
and treat θ as a *random variable*

Bayesian Estimation (IV)

- Bayesian estimation attempts to find the most probable values of the parameters, conditioned on the available data. Thus, parameters are regarded as random variables and our uncertainty in the values of the parameters is represented by a probability density function.
  - Before observation: a prior probability $p(\theta)$ describes the parameters
  - As we observe data:

    prior probability, $p(\theta) \rightarrow$ posterior probability, $p(\theta|D)$
- ***Bayesian Learning***: since some values of the parameters are more consistent with the data than others, we find that the posterior distribution is narrower than the prior distribution

Bayesian Learning



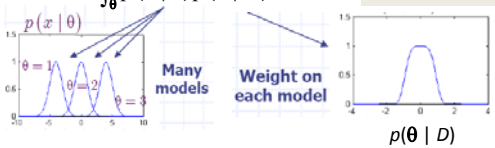
Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation.

Bayesian Estimation (V)

- Assume we have a prior $p(\boldsymbol{\theta})$ over models
- We have data (can assume i.i.d.) D
- Want to get a model to compute $p(\mathbf{x}|D)$… How to proceed?

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\theta}} p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$
$$= \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}, D) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$
$$= \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

average densities $p(\mathbf{x}|\boldsymbol{\theta})$ for ALL possible values of $\boldsymbol{\theta}$ weighted by its posterior probability



$p(\boldsymbol{\theta} \mid D)$

Bayesian Estimation (VI)

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

- If $p(\boldsymbol{\theta}|D)$ is sharply peaked at $\boldsymbol{\theta}_p \rightarrow p(\mathbf{x} \mid D) \approx p(\mathbf{x} \mid \boldsymbol{\theta}_p)$
- Posterior distribution for $\boldsymbol{\theta}$

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

- The equation shows how the observation of a set of training samples is incorporated to convert prior probability to posterior probability
- *Bayesian method does not commit to a particular value of $\boldsymbol{\theta}$, but uses the entire distribution*

Bayesian Estimation Summary

$$P(\omega_i \mid \mathbf{x}) = P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D_i) P(\omega_i \mid D)}{p(\mathbf{x} \mid D)}$$

$$\int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

$$\frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(D)}$$

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

BE: Univariate Gaussian (I)

- Univariate Gaussian case: $p(\mu|D)$

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

$p(\mu) \sim N(\mu_0, \sigma_0^2)$ prior density of the parameter

- Assume that
  - μ is the only unknown parameter
  - σ^2, μ_0 and σ_0^2 are known
- Need to find $p(\mu|D)$

$$p(\mu \mid D) = \frac{p(D \mid \mu)p(\mu)}{\int p(D \mid \mu)p(\mu)d\mu} = \alpha \prod_{k=1}^{n} p(x_k \mid \mu)p(\mu)$$

BE: Univariate Gaussian (II)

$$p(\mu \mid D) = \alpha p(\mu) \prod_{k=1}^{n} p(x_k \mid \mu)$$

$$= \alpha \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2 + \sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \quad (1)$$

This is a Gaussian

Let $p(\mu \mid D) \sim N(\mu_n, \sigma_n^2)$ i.e

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] \quad (2)$$

BE: Univariate Gaussian (III)

- Identifying coefficients in (1) and (2) yields:

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \quad \text{and} \quad \frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}$$

$$= \frac{n\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2} \qquad = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}$$

where $\hat{\mu}_n$ is the sample mean

- Solve the equations for μ_n and σ_n^2. Then

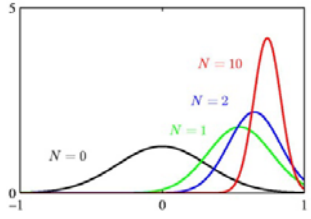
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

BE: Univariate Gaussian (IV)

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \quad \text{and} \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- If $n \to \infty$, then $\sigma_n^2 \to 0$. That is, additional observation decreases the uncertainty about μ_n, our best guess for μ, and eventually $p(\mu \mid D)$ becomes sharply peaked at μ_n (Bayesian learning).



BE: Univariate Gaussian (V)

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \quad \text{and} \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

- μ_n : linear combination of $\hat{\mu}_n$ and μ_0
  - If $\sigma_0^2 \neq 0$ then $\mu_n \to \hat{\mu}_n$ as $n \to \infty$
  - If $\sigma_0^2 = 0$ then $\mu_n = \mu_0$: very strong prior for $\mu = \mu_0$, observation cannot change it
  - If $\sigma_0^2 >> \sigma^2$ then $\mu_n \to \hat{\mu}_n$: uncertainty about the prior is large so does not have much influence (our guess for the mean becomes the sample mean
- If $\sigma^2/\sigma_0^2 \neq \infty$, then μ_n will converge to the sample mean after enough samples are taken

BE: Univariate Gaussian (VI)

- Class-conditional density for $p(x|D)$
  - $p(\mu|D)$ computed
  - $p(x|D)$ remains to be computed!

$$p(x \mid D) = \int p(x \mid \mu)\, p(\mu \mid D)\, d\mu = \int N(\mu, \sigma^2) N(\mu_n, \sigma_n^2)\, d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma, \sigma_n)$$

i.e. $p(x|D)$ is normally distributed and $p(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$

(Desired class-conditional density $p(x|D_i, \omega_i)$)

- $p(x|D_i, \omega_i)$ together with $P(\omega_i)$, and using Bayes formula, we obtain the posterior probabilities $P(\omega_i|x, D)$

BE: Multivariate Gaussian

- Direct generalization of the univariate case
- Read Section 3.4.3 of the textbook

BE: General Theory (I)

- $p(\mathbf{x}|D)$ computation can be applied to any situation in which the unknown density can be parameterized. The basic assumptions are:
  - Known form of $p(\mathbf{x}|\boldsymbol{\theta})$, but unknown value of $\boldsymbol{\theta}$
  - Known prior density $p(\boldsymbol{\theta})$:
    contains our knowledge about $\boldsymbol{\theta}$
  - The rest of our knowledge $\boldsymbol{\theta}$ is contained in a set D of n random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ that follows unknown $p(\mathbf{x})$

BE: General Theory (II)

- Basic problem:
  "Compute the posterior density $p(\boldsymbol{\theta}|D)$" and then
  "Derive $p(\mathbf{x}|D)$"

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

- Using Bayes formula, we have:

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

and by independence assumption: $p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$

Maximum-a-Posteriori vs. Maximum-likelihood

- Integration is difficult and rarely solvable, so approximate it…

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(D)} d\boldsymbol{\theta}$$

$$\approx \int_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) d\boldsymbol{\theta} = p(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$$

where

$$\hat{\boldsymbol{\theta}} = \begin{cases} \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) & \text{MAP} \\ \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}) & \text{ML (uniform } p(\boldsymbol{\theta})) \end{cases}$$

MAP vs. ML (cont'd)

- **Maximum a posteriori (MAP)** estimates: maximize the log of the posterior estimate with respect to parameters $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \left[\ln p(D \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \right]$$

- MAP is like ML, but also has penalty function on different models… $p(\boldsymbol{\theta})$ favors some over others. (If the prior distribution is flat (i.e. independent of $\boldsymbol{\theta}$), then the MAP estimate matches the ML estimate)

Recursive Bayes

- Given a training set of n samples $D^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

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

$$\Rightarrow p(D^n \mid \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta})$$

$$\Rightarrow p(\boldsymbol{\theta} \mid D^n) = \frac{p(D^n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D^n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$= \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D^{n-1})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D^{n-1}) d\boldsymbol{\theta}} \qquad (1)$$

Posterior is used as a prior in the next step

- **Recursive Bayes** approach to parameter estimation
  - Start with $p(\boldsymbol{\theta}|D^0) = p(\boldsymbol{\theta})$
  - Apply eq.(1): produces the sequence of densities $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}|\mathbf{x}_1)$, $p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$,… (*incremental* or *on-line learning*)

See Example 1 in page 98.

ML vs. Bayes Parameter Estimation

	ML	Bayes
parameters	Fixed value	Random variables
assumptions	It avoids assumptions on prior information and it is analytically easier to solve although some estimates can be biased	It permits including a priori information about the unknown, but the analytical derivation are cumbersome
computational complexity	Low (differential calculus)	possibly complex multidimensional integration
accuracy		More accurate statistically because it considers the uncertainty in estimating the parameters.
size of training set	Estimates do not change significantly as the training set size increases.	Estimates continuously improves as the training set size increases.
	For ordinary cases, both approaches give similar results with sufficient sample data	
interpretability	Easier (single best solution)	Complicated and harder (weighted average of models (parameters))
Confidence in the prior information	Solution must be of the assumed parametric form (e.g. Example 1)	Solution may not be of the parametric form originally assumed (e.g. Example 1).

partly from S. Iliescu