CS 7616 Pattern Recognition

Bayes Normals and more...

Aaron Bobick
School of Interactive Computing
Administrivia

- Slight change in course material assumptions:
  - Readings will now be assigned from Murphy (see class web site)
  - The PDF is available through the library (see web site)
    - But you should probably buy the book.

- First problem set will hopefully be out by Tues. Due in one week, Jan 28\(^{th}\), 11:55pm.
  - General description: for a trio of data sets (one common, one from the sets we provide, one from those sets or your own), use parametric density estimation for normal densities to find best result. Use both MLE methods and Bayes.
Outline for “today”

- Minimum loss decision function for Normal distributions
  - Introduces linear classifiers
  - Generally quadratic
- Where do we get Normals from – Parametric Density estimation
  - Maximum likelihood
  - MAP
    - Bayesian – priors, dogma
      - Using Bayesian distribution for classification
- Why and how normals?
  - What everyone knows
  - Maximum entropy
  - Minimum KL divergence
Recall: A special loss function

- Cost $\lambda_{ij}$ is 0 if $i = j$, 1 otherwise.Called zero-one loss function (duh).

- Which gives a ratio test: choose $\alpha_1$ if

$$\frac{p(x | \omega_1)}{p(x | \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

- Therefore need to estimate $p(x | \omega_i)\ldots$
Parameter Estimation

- Bayesian Decision Theory allows us to design an optimal classifier if we know the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(x/\omega_i)$.

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

- Estimate $P(\omega_i)$ and $p(x \mid \omega_i)$ from training examples.
  - Estimating $P(\omega_i)$ is usually not difficult. (Why?)
  - Estimating $p(x/\omega_i)$ is more difficult! (Why?)
    - Number of samples is often too small
    - Dimensionality of feature space is large.
Parameter Estimation (cont’d)

- Assumptions
  - We are given a set of training samples $D = \{x_1, x_2, \ldots, x_n\}$, where the samples were drawn according to $p(x/\omega_j)$
  - $p(x/\omega_j)$ has known parametric form, e.g.,
    $$p(x/\omega_i) \sim N(\mu_i, \Sigma_i) \quad \Theta=(\mu_i, \Sigma_i)$$

- Parameter estimation problem:
  Given $D$, find the best possible $\Theta$
Main Methods in Parameter Estimation

- Maximum Likelihood (ML)
  - ML estimation is usually simpler than alternative methods.
  - More accurate estimates as the number of training samples increases.
  - If the model chosen for $p(x/\theta)$ is correct, and independence assumptions among variables are true, ML will give very good results.
  - Otherwise, ML will give “poor” results. Especially when “small” data

- Bayesian Estimation (BE)
Bayesian Estimation (BE)

- Assumes that the parameters $\theta$ are random variables that have some known a-priori distribution $p(\theta)$.
- Using the training examples $D$, BE converts $p(\theta)$ to $p(\theta|D)$.
- BE estimates a distribution rather than making point estimates like ML or MAP.

$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

Note: BE solution might not be of the parametric form assumed.
The Role of Training Examples

- If \( p(x | \omega_i) \) and \( P(\omega_i) \) are known, Bayes’ rule allows us to compute the posterior probabilities \( P(\omega_i | x) \):

  \[
P(\omega_i | x) = \frac{p(x | \omega_i)P(\omega_i)}{\sum_j p(x | \omega_j)P(\omega_j)}
  \]

- The role of the training examples \( D_i \) can be emphasized by using them in the computation of the posterior probabilities:

  \[
P(\omega_i | x, D_i)
  \]
The Role of Training Examples (cont’d)

\[
P(\omega_i | x, D_i) = \frac{p(x, D_i | \omega_i) P(\omega_i)}{p(x, D_i)} = \frac{p(x | D_i, \omega_i) p(D_i | \omega_i) P(\omega_i)}{p(x / D_i) p(D_i)} = \\
= \frac{p(x | \omega_i, D_i) P(\omega_i | D_i)}{p(x | D_i)} = \frac{p(x | \omega_i, D_i) P(\omega_i | D_i)}{\sum_j p(x, \omega_j | D_j)} = \\
= \frac{p(x | \omega_i, D_i) P(\omega_i | D_i)}{\sum_j p(x | \omega_j, D_j) P(\omega_j | D_j)}
\]
The Role of Training Examples (cont’d)

• This implies that the training examples $D_i$ can help us to determine both the class-conditional densities and the prior probabilities:

$$P(\omega_i \mid x, D_i) = \frac{p(x \mid \omega_i, D_i)P(\omega_i \mid D_i)}{\sum_j p(x \mid \omega_j, D_j)P(\omega_j \mid D_j)}$$

• To simplify things, let’s assume that $P(\omega_i / D) = P(\omega_i)$:

$$P(\omega_i \mid x, D_i) = \frac{p(x \mid \omega_i, D_i)P(\omega_i)}{\sum_j p(x \mid \omega_j, D_j)P(\omega_j)}$$
Bayesian Estimation

- Need to estimate $p(x | \omega_i, D_i)$ for every class $\omega_i$

- If the samples in $D_j$ give no information about $\theta_i$ ($i \neq j$) then we need to solve $c$ independent problems of:

  "given $D$, estimate $p(x/D)$"
BE Approach

• Estimate $p(x/D)$ as follows:

$$p(x \mid D) = \int p(x, \theta \mid D) d\theta = \int p(x \mid \theta, D) p(\theta \mid D)d\theta$$

• Since the distribution is known completely given $\theta$, we have:

$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D)d\theta$$

• Important equation: links $p(x/D)$ with $p(\theta/D)$ \textit{which is a distribution}
BE Main Steps

(1) Compute $p(\theta / D)$:

$$p(\theta / D) = \frac{p(D / \theta)p(\theta)}{p(D)} = a \prod_{k=1}^{n} p(x_k / \theta) p(\theta)$$

(2) Compute $p(x / D)$:

$$p(x / D) = \int p(x / \theta) p(\theta / D) d\theta$$

$p(x/D)$ will replace $p(x/\omega_i, D_i)$ in Bayes’ rule:

$$P(\omega_i / x, D_i) = \frac{p(x / \omega_i, D_i)P(\omega_i)}{\sum_j p(x / \omega_j, D_j)P(\omega_j)}$$
Relation to ML Solution

\[ p(x / D) = \int p(x / \theta) p(\theta / D) d\theta \]

- If we are less certain about the exact value of \( \theta \), we should consider a weighted average of \( p(x / \theta) \) over the possible values of \( \theta \).

Bayesian estimation approach estimates a distribution for \( p(x/D) \) rather than making point estimates like ML.
Relation to ML Solution (cont’d)

- Suppose \( p(\theta / D) \) peaks very sharply at \( \theta = \hat{\theta} \), and \( p(\hat{\theta}) \neq 0 \), then \( p(x/D) \) can be approximated as follows:

\[
p(x / D) \approx p(x / \hat{\theta})
\]

(i.e., the best estimate is obtained by setting \( \theta = \hat{\theta} \))

- This is the ML solution (i.e., \( p(D/\theta) \) peaks at \( \hat{\theta} \) too)

since \( p(\theta / D) = \frac{p(D / \theta)p(\theta)}{p(D)} \)
Case 1: Univariate Gaussian, Unknown $\mu$

- Assume:

$$p(x|\mu) \sim N(\mu, \sigma^2) \text{ and } p(\mu) \sim N(\mu_0, \sigma_0^2)$$

$D = \{x_1, x_2, \ldots x_n\}$ are independently drawn

- Now compute:

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

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

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

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

(30)
Case 1: Univariate Gaussian, Unknown $\mu$ (cont’d)

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

\[ \mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \]

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

\[ \bar{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k \]

- $\mu_n$ is our best guess for $\mu$; $\sigma_n^2$ measures the uncertainty.
Case 1: Univariate Gaussian, Unknown $\mu$ (cont’d)

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

- $\mu_n$ is a linear combination of $\bar{x}_n$ and $\mu_0$ (always lies between)
- If $\sigma_0 \neq 0$ then $\mu_n$ approaches $\bar{x}_n$ as $n \to \infty$ (ML estimate)
- If $\sigma_0 = 0$, then $\mu = \mu_0$; If $\sigma_0 \gg \sigma$ then $\mu_n = \bar{x}_n$ (ML estimate)

$n \to \infty$ implies more samples!
Case 1: Univariate Gaussian, Unknown $\mu$ (cont’d)

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

- $\sigma_n^2$ approaches $\sigma^2/n$ as $n$ increases (more observations will decrease our uncertainty about $\mu$)

- $p(\mu/D)$ becomes sharply peaked (*Bayesian Learning*) as $n$ increases.
Case 1: Univariate Gaussian, Unknown $\mu$ (cont’d)

**Bayesian Learning**

*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.
Case 1: Univariate Gaussian, Unknown $\mu$
(cont’d)

Compute $p(x/D)$:

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

or $p(x/D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$

As the number of samples increases, $p(x/D)$ converges to $p(x/\mu)$
Case 2: Multivariate Gaussian, Unknown $\mu$

Assume $p(x/\mu) \sim N(\mu, \Sigma)$ and $p(\mu) \sim N(\mu_0, \Sigma_0)$

$D = \{x_1, x_2, \ldots, x_n\}$ (independently drawn)

Compute $p(\mu/D)$:

$$p(\mu \mid D) = \frac{p(D \mid \mu)p(\mu)}{p(D)} \propto \prod_{k=1}^{n} p(x_k \mid \mu)p(\mu)$$
Case 2: Multivariate Gaussian, Unknown $\mu$
(cont’d)

- Substituting the expressions for $p(x_k/\mu)$ and $p(\mu)$:

$$p(\mu / D) = c \times \exp[-\frac{1}{2}(\mu - \mu_n)^t \Sigma_n^{-1}(\mu - \mu_n)]$$

where

$$\mu_n = \Sigma_0 (\Sigma_0 + \frac{1}{n} \Sigma)^{-1} \bar{x}_n + \frac{1}{n} \Sigma (\Sigma_0 + \frac{1}{n} \Sigma)^{-1} \mu_0$$

$$\Sigma_n = \Sigma_0 (\Sigma_0 + \frac{1}{n} \Sigma)^{-1} \frac{1}{n} \Sigma$$

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$$
Case 2: Multivariate Gaussian (cont’d)

Compute $p(x/D)$:

$$p(x \mid D) = \int p(x \mid \mu)p(\mu \mid D)d\mu \sim N(\mu_n, \Sigma + \Sigma_n)$$

And now you can do that Bayesian thing….

$$P(\omega_i \mid x, D_i) = \frac{p(x \mid \omega_i, D_i)P(\omega_i)}{\sum_j p(x \mid \omega_j, D_j)P(\omega_j)} \propto p(x \mid D_i) P(\omega_i)$$
Now some normal properties

- Everyone knows central limit theorem
- But not everyone knows
  - Gaussian is a minimal assumption (maximum entropy)
  - Best Gaussian to use is the obvious one...

- But first:
  - Entropy
  - KL Convergence
Entropy

- **Entropy** measures the “uncertainty” of a density/distribution.
- Discrete case:
  \[ H(p_1, p_2, \ldots, p_n) = -\sum_{i} p_i \log(p_i) \]
  - Entropy is maximum when \( p_i = p \)
  - Example for n=2 consider (.25, .75) and (.5, .5)
    - \(-.25*(\log(1)-\log(4))- .75*(\log(3)-\log(4)) = -.25*(-2)-.75(\log(3)= .81)
    - \(.5*(\log(1)-\log(2))*2 = 1.0\)
  
- Continuous case
  \[ h(p(x)) = -\int p(x) \log(p(x)) dx \]
- For Gaussian \( g(x) \sim N(\mu, \sigma^2) \):
  \[ h(g) = \frac{1}{2} \log(2\pi e\sigma^2) \]
KL divergence

- KL divergence is a measure of the disagreement between two distributions/densities
- Discrete: 
  \[ D_{KL}(P||Q) = \sum_i \ln \left( \frac{P(i)}{Q(i)} \right) P(i) \]
- Continuous: 
  \[ D_{KL}(P||Q) = \int_{-\infty}^{\infty} \ln \left( \frac{p(x)}{q(x)} \right) p(x) \, dx \]

- Properties:
  - KL is not symmetric \( D_{KL}(P||Q) \neq D_{KL}(Q||P) \) so not a “distance”
  - \( D_{KL}(P||Q) \geq 0 \) (proof based on \( \ln(x) < x - 1 \))
  - \( D_{KL}(P||Q) = 0 \) iff the two densities/distributions are the same
Gaussian is maximum entropy

- Suppose we have a density \( f(x) \) that has mean \( \mu \) and variance \( \sigma^2 \) and that Gaussian \( g(x) \) has same mean and variance

\[
0 \leq D_{KL}(f || g) = \int_{-\infty}^{\infty} f(x) \log \left( \frac{f(x)}{g(x)} \right) dx = -h(f) - \int_{-\infty}^{\infty} f(x) \log(g(x)) dx.
\]

- Right hand term:

\[
\int_{-\infty}^{\infty} f(x) \log(g(x)) dx = \int_{-\infty}^{\infty} f(x) \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx
\]

\[
= \int_{-\infty}^{\infty} f(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx + \log(e) \int_{-\infty}^{\infty} f(x) \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) dx
\]

\[
= -\frac{1}{2} \log(2\pi\sigma^2) - \log(e) \frac{\sigma^2}{2\sigma^2}
\]

\[
= -\frac{1}{2} \log(2\pi\sigma^2) + \log(e)
\]

\[
= -\frac{1}{2} \log(2\pi e\sigma^2)
\]
Continuing...

From last slide:

\[ 0 \leq D_{KL}(f\|g) = \int_{-\infty}^{\infty} f(x) \log \left( \frac{f(x)}{g(x)} \right) \, dx = -h(f) - \int_{-\infty}^{\infty} f(x) \log(g(x)) \, dx. \]

\[ 0 \leq D_{KL}(f \| g) = -h(f) - (-h(g)) \]

So:

\[ h(g) \geq h(f) \]

- What does this say about Gaussians and why does it matter for pattern recognition?
Related: “closest Gaussian”

We assume $p_1(x) \equiv p(x|\omega_1) \sim N(\mu_1, \Sigma_1)$ but that $p_2(x) \equiv p(x|\omega_1)$ is arbitrary.

- So KL (using that $p_1$ is Gaussian):

$$D_{KL}(p_2, p_1) = \int p_2(x) \ln p_2(x) dx + \frac{1}{2} \int p_2(x) \left[ \ln(2\pi) + \ln|\Sigma| + (x - \mu)^t \Sigma^{-1} (x - \mu) \right] dx,$$

- Now find $\mu$ and $\Sigma$ that minimize the divergence:

$$\frac{\partial}{\partial \mu} D_{KL}(p_2, p_1) = - \int \Sigma^{-1} (x - \mu) p_2(x) dx = 0$$

$$\int p_2(x) (x - \mu) dx = \mathcal{E}_2[x - \mu] = 0$$

- So this implies $\mu$ for the approximating Gaussian should be the mean of the second distribution
Related: “closest Gaussian” (cont)

\[ D_{KL}(p_2, p_1) = \int p_2(x) \ln p_2(x) dx + \frac{1}{2} \int p_2(x) \left[ d\ln(2\pi) + \ln|\Sigma| + (x - \mu)^t \Sigma^{-1} (x - \mu) \right] dx, \]

- Nor for \( \Sigma \):

\[ \frac{\partial}{\partial \Sigma} D_{KL}(p_2, p_1) = 0 = \int p_2 \left[ -\Sigma^{-1} + (x - \mu)(x - \mu)^t \right] dx \]

(because \( \frac{\partial |A|}{\partial A} = |A|A^{-1} \) when \( A \) is symmetric)

- So: \( \mathcal{E}_2 \left[ \Sigma - (x - \mu)(x - \mu)^t \right] \) or \( \mathcal{E}_2 \left[ (x - \mu)(x - \mu)^t \right] = \Sigma. \)

- This implies the covariance of the 2\textsuperscript{nd} distribution should be the same as the Gaussian as well.