KHU-BME
MLPR
Lecture 13

Bayesian Estimation (DHS 3.3)

Bayesian Classifier (DHS 3.3.1)

$P(S_k|\underline{x}) \geq P(S_j|\underline{x})$ for all j=1 to k => $\underline{x} \in S_k$

$$P(S_k|\underline{x}) \propto P(\underline{x}|S_k)P(S_k)$$

What do we do if $P(S_j)$ and $P(\underline{x}|S_j)$ are unknown? Compute them using all information we have.

Given $\underline{z}$, the set of samples, compute the posterior probabilities $P(S_k|\underline{x},\underline{z})$ ← Final Goal

(i.e., use the training samples to compute the class-conditional density and prior density)

From Bayes theorm, $P(S_k|\underline{x},\underline{z}) = p(\underline{x}|S_k,\underline{z})P(S_k|\underline{z}) \,/\, \sum_{j=1}^{K} p(x|S_j,\underline{z})\, P(S_j|\underline{z})$

Assume $P(S_j|\underline{z}) = P(S_j)$ and $P(S_j)$ are known

Subdivide $\underline{z}$: $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_k$

where $\underline{z}_i$ contains all prototypes in class S_i

Assume $p(\underline{x}|S_i,\underline{z}) = p(\underline{x}|S_i,\underline{z}_i)$

$P(S_k|\underline{x},\underline{z}) = p(\underline{x}|S_k, \underline{z}_k)\, P(S_k) \,/\, \sum_{j=1}^{K} p(\underline{x}|S_j,\underline{z}_j)\, P(S_j)$

That is treat each class separately,

$P(\underline{x}|S_k)$ has known parametric form => $p(\underline{x}|S_k, \underline{\theta})$ is known.

The goal is to determine the likelihood, $p(x|S_k, \underline{z}_k)$ using the prototypes $\underline{z}_k$. ← Goal

Parameter Distribution (DHS 3.3.2)

Now, our goal is to compute $p(x|\underline{z}_k)$, which can be computed from
$p(x|\underline{z}_k)=\int p(x,\underline{\theta}|\underline{z}_k)d\underline{\theta}$
since the selection of x and the selection of the training samples in $\underline{z}_k$ is done independently, rewrite this
$p(x|\underline{z}_k)=\int p(x|\underline{\theta})p(\underline{\theta}|\underline{z}_k)d\underline{\theta}$ (*)
This equation links the class-conditional density $p(x|\underline{z}_k)$ to the posterior density $p(\underline{\theta}|\underline{z}_k)$ for the unknown parameter vector.

If $p(\underline{\theta}|\underline{z}_k)$ peaks very sharply about the some value $\hat{\underline{\theta}}$, we obtain $p(x|\underline{z}_k) \approx p(x|\hat{\underline{\theta}})$.

This means the result is obtained by substituting $\hat{\underline{\theta}}$ (the estimate) for the true parameter.

Bayesian Parameter Estimation: Gaussian Case (DHS 3.4)

Goal: compute **(Goal 1)** $p(\underline{\theta}|\underline{z})$ and then **(Goal 2)** $p(\underline{x}|\underline{z})$ in (*). ← Two Goals

(Goal 1) Learning the mean of a normal density: get $p(\mu|z_k)$ (DHS 3.4.1): Univariate Case

Assumption: $p(x|\mu) = N(x,\mu,\sigma^2)$
μ=unknown
σ^2=known
$\theta=\mu$

Objective: determine $p(\mu|z_k)$
Assume: $p(\mu)=N(\mu,\mu_o,\sigma_o^2)$ known
μ_o=known
= a priori guess of parameter μ
σ_o^2=known
= variance or uncertainty of the guess μ_o

Bayes theorem $p(\mu|\underline{z})=[\,p(z|\mu)p(\mu)\,] / [\int p(z|\mu)\,p(\mu)\,d\mu\,]$

Denominator = $p(z) = 1/\alpha$

$\{z_k\} = \{x_1, x_2, \ldots, x_J\}$
= set of prototypes from same class k, independently drawn from the population

Drop k -> z & Drop S_k

x_i's independently drawn => $p(z|\mu) = \prod_{j=1}^{J} p(x_j|\mu)$

$$p(\mu|z) = \alpha \prod_{j=1}^{J} p(x_j|\mu)p(\mu)$$

$$= \alpha\, p(\mu) \prod_{j=1}^{J} p(x_j|\mu)$$

$p(\mu)=N(\mu,\mu_o, \sigma_o^2)$ (known)

$p(x_j|\mu)=N(x_j,\mu,\sigma^2)$ (μ unknown and σ^2 known)

$$p(\mu|\underline{z})=\alpha \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\} \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma}\right)^2\right\}$$

$$=\alpha' \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\} \exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\left(\frac{x_j-\mu}{\sigma}\right)^2\right\}$$

$$=\alpha' \exp\left\{-\frac{1}{2{\sigma_0}^2}\mu^2 + \frac{\mu_0}{{\sigma_0}^2}\mu - \frac{1}{2}\frac{\mu_0}{{\sigma_0}^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J} x_j^2 + \frac{\mu}{\sigma^2}\sum_{j=1}^{J} x_j - \frac{1}{2\sigma^2} J\mu^2\right\}$$

$$=\alpha'' \exp\left\{-\frac{1}{2}[(\frac{1}{{\sigma_0}^2}+\frac{J}{\sigma^2})\mu^2 - 2(\frac{\mu_0}{{\sigma_0}^2}+\frac{Jm_J}{\sigma^2})\mu]\right\} \qquad (**)$$

where $m_J = \frac{1}{J}\sum_{j=1}^{J} x_j$

Since p(μ|z)=normal

Now set, p(μ|z) = $\frac{1}{\sqrt{2\pi}\sigma_J}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_J}{\sigma_J}\right)^2\right\}$ = N(μ,μJ,σJ²)

=β $\exp\left\{-\frac{1}{2}[\frac{1}{\sigma_J{}^2}\mu^2-\frac{2\mu_J}{\sigma_J{}^2}\mu+\frac{\mu_J{}^2}{\sigma_J{}^2}]\right\}$ and compare to (**)

=> $\frac{1}{\sigma_J^2}=\frac{1}{\sigma_0^2}+\frac{J}{\sigma^2}$

$\frac{\mu_J}{\sigma_J^2}=\frac{\mu_0}{\sigma_0^2}+\frac{Jm_J}{\sigma^2}$

So $\boxed{\sigma_J^2=\frac{\sigma^2\sigma_0^2}{\sigma^2+J\sigma_0^2}}$ $\boxed{\mu_J=\sigma_J^2[\frac{\mu_0}{\sigma_0^2}+\frac{Jm_J}{\sigma^2}]}$

Therefore p(μ|z)=N(μ,μJ,σJ²)

$$\boxed{\begin{aligned}\mu_J&=\frac{\sigma^2}{\sigma^2+J\sigma_0{}^2}\mu_0+\frac{J\sigma_0^2}{\sigma^2+J\sigma_0^2}m_J\\ \sigma_J^2&=\frac{\sigma^2\sigma_0^2}{\sigma^2+J\sigma_0^2},\quad m_J=\frac{1}{J}\sum_{j=1}^{J}x_j\end{aligned}}$$

These equations show how prior is combined with empirical information in samples to obtain a posteriori density p($\underline{\theta}$ | $\underline{z}$)

μ_J is best estimate of μ after J observations.
σ_J is the uncertainty in the estimate μ_J.
This behavior is called Bayesian Learning (note DHS Fig. 3.2)

(Goal 2) Now get p(x|z): again univariate case (Read DHS 3.4.2)

Now we know p(μ|z).

Need to obtain the class-conditional density, $p(x|z)=p(x|S_k,z_k)$

Recall (*) => $p(x|z)=\int p(x|\mu)p(\mu|z)d\mu$

$$p(x|z)=\int \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\frac{1}{\sqrt{2\pi}\sigma_J}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_J}{\sigma_J}\right)^2\right\}d\mu$$

$$p(x|z)=\frac{1}{2\pi\sigma\sigma_J}\exp\left\{-\frac{1}{2}\left(\frac{(x-\mu_J)^2}{\sigma^2+\sigma_J{}^2}\right)\right\}f(\sigma,\sigma_J)$$

where $f(\sigma,\sigma_J)=\int\exp\left\{-\frac{1}{2}\frac{\sigma^2+\sigma_J{}^2}{\sigma^2\sigma_J{}^2}\left(\mu-\frac{\sigma_J{}^2\mu+\sigma^2\mu_J}{\sigma^2+\sigma_J{}^2}\right)^2\right\}d\mu$

This makes p(x|z) as a function of x is proportional to $\exp\left\{-\frac{1}{2}\left(\frac{(x-\mu_J)^2}{\sigma^2+\sigma_J{}^2}\right)\right\}$

Therefore p(x|z) is normally distributed with mean μ_J and $\sigma^2+\sigma_J{}^2$:

∴ $p(x|z)=N(x,\mu_J,\sigma^2+\sigma_J{}^2)$

[Summary] Bayesian Estimation Procedure

1. Estimate μ_J and σ_J by the boxed formulas and substitute into the equation for p(μ|z).
2. Find p(x|z) from above.
3. Use the following to find, $P(S_k|\underline{x},\underline{z})$, given p(x|z)=p(x|S,z) and p(S)

$$p(S_k|\underline{x},\underline{z}) = p(\underline{x}|S_k, \underline{z}_k)\, p(S_k) \Big/ \sum_{j=1}^{K} p(\underline{x}|S_j,\underline{z}_j)\, p(S_j)$$

(for class S_k, $p(x|z)=p(x|S_k,z_k)$)

Multivariate Extension

$p(x|\mu)=N(x,\mu,\Sigma)$

$p(\mu)=N(\mu,\mu_0, \Sigma_0)$

$p(x|z)=N(x,\mu_J, \Sigma+\Sigma_J)$

μ_0, Σ_0, Σ are known. μ is unknown

where

$$\boxed{\begin{aligned}&\mu_J = \Sigma_0[\Sigma_0+\Sigma/J]^{-1}m_J + (1/J)\,\Sigma[\Sigma_0 + (1/J)\,\Sigma]^{-1}\,\mu_0\\ &\Sigma_J=\Sigma_0[\Sigma_0+1/J)\,\Sigma]^{-1}\,(\Sigma/J)\end{aligned}}$$

$$m_J = \frac{1}{J}\sum_{j=1}^{J} x_j$$

μ_0=initial guess of μ

Σ_0=initial uncertainty.

Generalize the above technique => <u>General Bayesian Learning</u>

(Again, do for each class S_k separately)

1. General form of $p(\underline{x}|\underline{\theta})$ is known, but θ is not known exactly.
2. Initial knowledge about $\underline{\theta}$ is available as $p(\underline{\theta})$. Rest of our knowledge is a set of samples $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_J\} = \underline{z}$ of known classification drawn from a population of unknown density $p(\underline{x})$. ($\underline{x}_j$ is independent of $\underline{x}_k$)

Objective: Compute $p(\underline{x}|\underline{z})$ to obtain $P(S_k|\underline{x},\underline{z})$

Procedure

1. $p(\underline{x}|\underline{z})=\int p(\underline{x}|\underline{\theta})p(\underline{\theta}|\underline{z})d\underline{\theta}$ (*)

where

2. $p(\underline{\theta}|\underline{z}) = p(\underline{z}|\underline{\theta})p(\underline{\theta}) / \int p(\underline{z}|\underline{\theta})p(\underline{\theta})d\underline{\theta}$

where

3. $p(\underline{z}|\underline{\theta}) = \prod_{j=1}^{J} p(\underline{x}_j|\underline{\theta})$

Note: If $p(z|\theta)$ peaks at $\underline{\theta}=\underline{\hat{\theta}}$, then $p(\underline{\theta}|\underline{z})$ will peak at $\underline{\theta}=\underline{\hat{\theta}}$

ML: θ that maximizes $p(\underline{z}|\underline{\theta})$.

Recursive Bayesian Learning

From 3, $p(z^{(J)}|\theta) = p(x_J|\theta)\prod_{j=1}^{J-1} p(x_j|\theta)$

or $p(z^{(J)}|\theta) = p(x_J|\theta)\, p(z^{(J-1)}|\theta)$

From 2, $p(\theta|z^{(J)}) = \dfrac{[p(x_J\,|\,\theta)p(z^{(J-1)}\,|\,\theta)]p(\theta)}{\int[p(x_J\,|\,\theta)p(z^{(J-1)}\,|\,\theta)]p(\theta)d\theta}$

Use Bayes rule

$$p(z^{(J-1)}\,|\,\theta)p(\theta) = p(\theta\,|\,z^{(J-1)})p(z^{(J-1)})$$

$$p(\theta\,|\,z^{(J)}) = \frac{p(x_J\,|\,\theta)p(\theta\,|\,z^{(J-1)})p(z^{(J-1)})}{\int p(x_J\,|\,\theta)p(\theta\,|\,z^{(J-1)})p(z^{(J-1)})d\theta}$$

$p(\theta|z^{(o)}) = p(\theta)$

Sequence: $p(\theta)$ (initial guess with no data)
$p(\theta|z^{(1)}) = p(\theta|x_1)$
$p(\theta|z^{(2)}) = p(\theta|x_1,x_2)$
… Usually converges to a delta function.

Which Method is Better? Maximum-Likelihood or Bayes Methods (DHS 3.5.1)

- Computational Complexity: ML is preferred since it requires merely differential calculus techniques or gradient search w.r.t. parameters. Bayes methods require complex multidimensional integration.
- ML will be easier to interpret and understand, but Bayesian gives a weighted average of models or parameters, leading to solutions more complicated and harder to understand.
- Bayesian uses more information brought into the problem than ML.
- If more reliable information available, Bayes gives better results.
- Bayesian with a flat prior gives same results as ML.
- Bayesian balances between the accuracy of the estimation and its variance.
- Three Error Sources
  - Bayes error: due to overlapping densities. Cannot be eliminated.
  - Model error: due to an incorrect model. With better models, can be reduced.
  - Estimation error: due to a finite sample. Reduced with more training data.

Problems of Dimensionality (DHS 3.7)

- How classification accuracy depends upon the dimensionality and the amount of training data
- The computational complexity of designing the classifier.
- For Bayes classifier, the most useful features are the ones that offer bigger differences between the means than the standard deviations, thus reducing the probability of error.
- An obvious way is to introduce new independent features.
- If performance of a classifier is poor, it is natural to utilize new features, particularly ones that will help separate the class pairs most frequently confused.
- But increasing the number of features increases the cost and complexity of both the feature extractor and the classifier.
- In general, the performance should improve
- See DHS Fig. 3.3

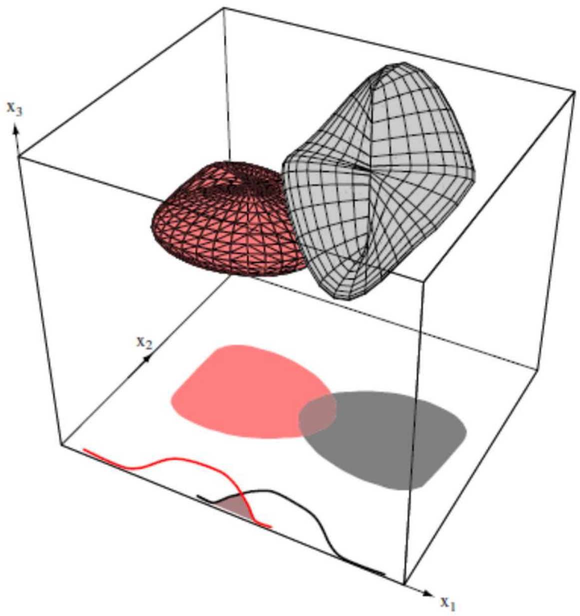

Figure 3.3: Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace — here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace — there can be greater overlap of the projected distributions, and hence greater Bayes errors.

Overfitting (DHS 3.7.3)

- See DHS Fig. 3.4

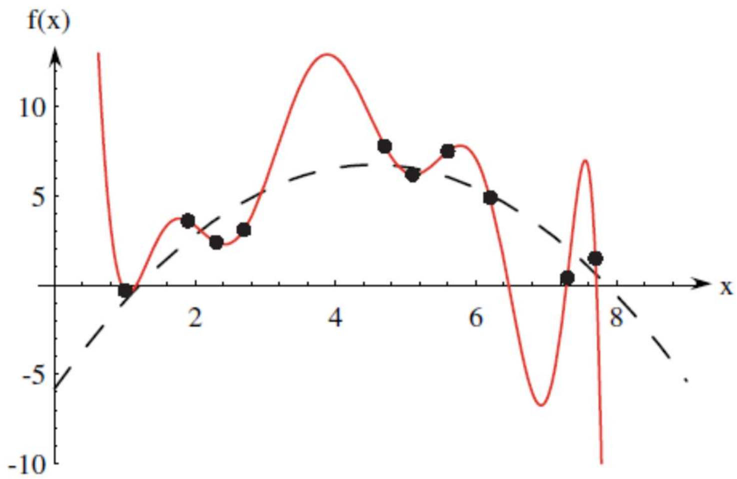

Figure 3.4: The "training data" (black dots) were selected from a quadradic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since it would lead to better predictions for new samples.

** Principle Component Analysis (DHS 3.8.1) and Fisher Linear Discriminant (DHS 3.8.2) will be covered later in Unsupervised Classification.*