Parametric Methods

Berlin Chen, 2005

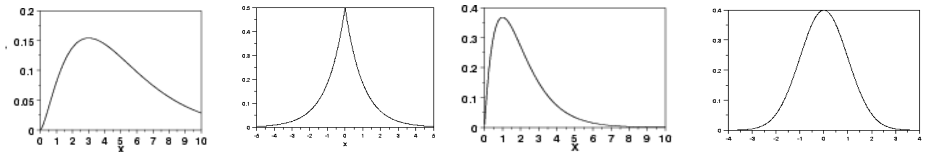
References:

1. *Introduction to Machine Learning* , Chapter 4

Introduction

- Statistic
  - Any value that is calculated from a given sample
  - Statistical inference: make a decision using the information provided by a sample (or samples)

- Parametric methods
  - Assume that samples are drawn from some distribution that obeys a known model $p(x)$



  - Advantage: the model is well defined up to a small number of parameters
    - E.g., mean and variance are sufficient statistics for the Gaussian distribution
  - Model parameters are typically estimated by either maximum likelihood estimation or Bayesian (MAP) estimation
  - Data samples are assumed to be univariate (scalar variables) here

Maximum Likelihood Estimation (MLE)

- Assume samples $X=\{x^1,x^2,\ldots,x^t,\ldots,x^N\}$ are independent and identically distributed (*iid*), and drawn from some known probability distribution
  - $x^t \sim p(x|\theta)$
  - θ : model parameters (assumed to be fixed but unknown here)

- MLE attempts to find θ that make X the most likely to be drawn
  - Namely, maximize the likelihood of samples

$x^1,\ldots,x^N$ are *iid*

$$l(\theta|X) = p(X|\theta) = p(x^1,\cdots,x^N|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

MLE (cont.)

- Because logarithm will not change the value of θ when it take its maximum
  - Finding θ that maximizes the likelihood of the samples is equivalent to finding θ that maximizes the log likelihood of the samples

$$L(\theta|X) = \log\ l(\theta|X) = \sum_{t=1}^{N} \log\ p(x^t|\theta)$$

$$a \geq b$$
$$\Rightarrow \log\ a \geq \log\ b$$

  - As we will see, logarithmic operation can further simplify the computation when estimating the parameters of those distributions that have exponents

MLE: Bernoulli Distribution

- Bernoulli Distribution
  - Random variable x takes either the value 1 (with probability r) or the value 0 (with probability $1-r$)
    - Can be thought of as x is generated form two distinct states
  - The associated probability distribution

$$P(x)=r^x(1-r)^{1-x} \quad , x\in\{0,1\}$$

- The log likelihood for a set of *iid* samples drawn from Bernoulli distribution

$$X=\{x^1,\ldots,x^t,\ldots,x^N\}$$

$$L(r|X)=\log\prod_{t=1}^{N} r^{(x^t)}(1-r)^{(1-x^t)}$$

$$\theta \quad = \left(\sum_{t=1}^{N} x^t\right)\log r + \left(N-\sum_{t=1}^{N} x^t\right)\log(1-r)$$

MLE: Bernoulli Distribution (cont.)

- MLE of the distribution parameter r

$$\hat{r} = \frac{\sum_{t=1}^{N} x^t}{N}$$

  - The estimate for r is the ratio of the number of occurrences of the event ($x^t = 1$) to the number of experiments

- The expected value for X

$$E[X] = \sum_{x \in \{0,1\}} x \cdot p(x) = 0 \cdot (1 - r) + 1 \cdot r = r$$

MLE: Bernoulli Distribution (cont.)

- Appendix A

$$\frac{\partial L(r|X)}{\partial r} = \frac{\partial\left[\left(\sum_{t=1}^{N} x^t\right)\log r + \left(N - \sum_{t=1}^{N} x^t\right)\log(1-r)\right]}{\partial r} = 0$$

$$\Rightarrow \frac{\left(\sum_{t=1}^{N} x^t\right)}{r} - \frac{\left(N - \sum_{t=1}^{N} x^t\right)}{1-r} = 0$$

$$\Rightarrow \hat{r} = \frac{\sum_{t=1}^{N} x^t}{N}$$

$$\frac{\partial \log y}{\partial y} = \frac{1}{y}$$

MLE: Multinomial Distribution

- Multinomial Distribution
  - A generalization of Bernoulli distribution
  - A value of a random event x can be one of *K* mutually exclusive and exhaustive states $s_i = \{s_1, s_2, \cdots, s_K\}$
  - The associated probability distribution

$$p(x) = \prod_{i=1}^{K} r_i^{s_i^t}, \qquad \sum_{i=1}^{K} r_i = 1$$

$$s_i^t = \begin{cases} 1 & \text{if } x \text{ choose state } s_i \\ 0 & \text{otherwise} \end{cases}$$

- The log likelihood for a set of *iid* samples drawn from Bernoulli distribution

$$L(\boldsymbol{r} | X) = \log \prod_{t=1}^{N} \prod_{i=1}^{K} r_i^{s_i^t} \qquad X = \{x^1, \ldots, x^t, \ldots, x^N\}$$

MLE: Multinomial Distribution (cont.)

- MLE of the distribution parameter r_i

$$\hat{r}_i = \frac{\sum_{t=1}^{N} s_i^t}{N}$$

  - The estimate for r_i is the ratio of the number of experiments with outcome of state i ($s_i^t = 1$) to the number of experiments

MLE: Multinomial Distribution (cont.)

- Appendix B

$$L(\boldsymbol{r}|X) = \log \prod_{t=1}^{N}\prod_{i=1}^{K} r_i^{s_i^t} = \sum_{t=1}^{N}\sum_{i=1}^{K} \log r_i^{s_i^t}, \quad \text{with constraint} \quad : \sum_{i=1}^{K} r_i = 1$$

$$\frac{\partial \overline{L}(\boldsymbol{r}|X)}{\partial r_i} = \frac{\partial\left[\sum_{t=1}^{N}\sum_{i=1}^{K} s_i^t \cdot \log r_i + \lambda\left(\sum_{i=1}^{K} r_i - 1\right)\right]}{\partial r_i} = 0$$

Lagrange Multiplier

$$\Rightarrow \sum_{t=1}^{N} s_i^t \cdot \frac{1}{r_i} + \lambda = 0$$

$$\Rightarrow r_i = -\frac{1}{\lambda}\sum_{t=1}^{N} s_i^t$$

$$\Rightarrow \sum_{i=1}^{K} r_i = 1 = -\frac{1}{\lambda}\sum_{t=1}^{N}\left(\sum_{i=1}^{K} s_i^t\right)$$

=1

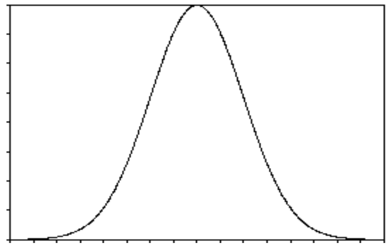
$$\Rightarrow \lambda = -N$$

$$\Rightarrow \hat{r}_i = \frac{\sum_{t=1}^{N} s_i^t}{N}$$

MLE: Gaussian Distribution

- Also called Normal Distribution
  - Characterized with mean μ and variance σ^2

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$



  - Recall that mean and variance are sufficient statistics for Gaussian
- The log likelihood for a set of *iid* samples drawn from Gaussian distribution

$$L(\mu, \sigma | X) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x^t-\mu)^2}{2\sigma^2}\right)} \qquad X = \{x^1, \ldots, x^t, \ldots, x^N\}$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}$$

MLE: Gaussian Distribution (cont.)

- MLE of the distribution parameters μ and σ^2

$$m = \hat{\mu} = \frac{\sum_{t=1}^{N} x^t}{N}$$

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{t=1}^{N} \left(x^t - m\right)^2}{N}$$

- Reminder that μ and σ^2 are still fixed but unknown

MLE: Gaussian Distribution (cont.)

- Appendix C

$$L(\mu,\sigma|X) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L(\mu,\sigma|X)}{\partial\mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu)^2 = 0 \Rightarrow \hat{\mu} = \frac{\sum_{t=1}^{N}x^t}{N}$$

$$\frac{\partial L(\mu,\sigma|X)}{\partial\sigma^2} = 0 \Rightarrow -N + \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{t=1}^{N}(x^t-\hat{\mu})^2}{N}$$

Evaluating an Estimator: Bias

- Suppose X is a sample from a population distribution with a parameter θ
- Let $d = d(X)$ be an estimator of θ and bias of the estimator d is defined as
  - The expected difference between d and θ

$$b_\theta(d) = E[d - \theta] = E[d] - \theta$$

- An unbiased estimator d has the property that

$$b_\theta(d) = E[d] - \theta = 0$$

- d is an asymptotically unbiased estimator
  - The bias goes to zero as the sample size $|X|$ goes to infinite

Evaluating an Estimator: Variance

- The variance measures how much, on average, the estimator d varies around the expected value $E[d]$

$$\mathrm{Var}(d) = E\left[(d - E[d])^2\right]$$

  – As we will see later: the smaller the sample size $|X|$, the larger the variance

- The mean square error of the estimator d is defined as

$$r(d, \theta) = E\left[(d - \theta)^2\right]$$

  – Measure how much the estimator d is different from θ

Evaluating an Estimator (cont.)

- The mean square error of the estimator d can be further decomposed into two parts respectively composed of bias and variance

$$
\begin{aligned}
r(d,\theta) &= E\left[(d-\theta)^2\right] \\
&= E\left[(d - E[d] + E[d] - \theta)^2\right] \\
&= E\left[(d - E[d])^2 + (E[d]-\theta)^2 + 2(d-E[d])(E[d]-\theta)\right] \\
&= E\left[(d - E[d])^2\right] + E\left[\underbrace{(E[d]-\theta)^2}_{\text{constant}}\right] + 2E\left[(d-E[d])\underbrace{(E[d]-\theta)}_{\text{constant}}\right] \\
&= E\left[(d - E[d])^2\right] + (E[d]-\theta)^2 + 2\underbrace{E\left[(d-E[d])\right]}_{0}(E[d]-\theta) \\
&= \underbrace{E\left[(d - E[d])^2\right]}_{\text{variance}} + \underbrace{(E[d]-\theta)^2}_{\text{bias}^2}
\end{aligned}
$$

Evaluating an Estimator (cont.)

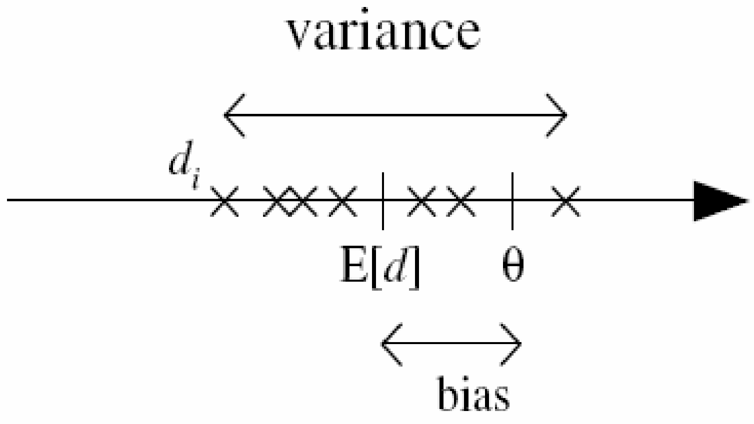


Figure 4.1: θ is the parameter to be estimated. d_i are several estimates (denoted by '$\times$') over different samples. Bias is the difference between the expected value of d and θ. Variance is how much d_i are scattered around the expected value. We would like both to be small.

Evaluating an Estimator (cont.)

- Example 1: **sample average and sample variance**
  - Assume samples $X=\{x^1,x^2,\ldots,x^t,\ldots,x^N\}$ are independent and identically distributed (*iid*), and drawn from some known probability distribution with mean μ and variance σ^2
    - Mean $\mu = E[x^t] = \sum_{x^t} x^t \cdot p(x^t)$
    - Variance $\sigma^2 = E\left[(x^t-\mu)^2\right] = E\left[(x^t)^2\right] - \left(E[x^t]\right)^2$
    - Sample average (mean) for the observed samples $m = \frac{1}{N}\sum_{t=1}^{N} x^t$
    - Sample variance for the observed samples $s^2 = \frac{1}{N}\sum_{t=1}^{N}(x^t - m)^2$

or $s^2 = \frac{1}{N-1}\sum_{t=1}^{N}(x^t-m)^2$?

Evaluating an Estimator (cont.)

- Example 1 (count.)
  - Sample average m is an unbiased estimator of the mean μ

$$E[m] = E\left[\frac{1}{N}\sum_{t=1}^{N} x^t\right] = \frac{1}{N}\sum_{t=1}^{N} E\left[x^t\right] = \frac{N \cdot \mu}{N} = \mu$$

- m is also a consistent estimator: $\mathrm{Var}(m) \to 0$ as $N \to \infty$

$$\mathrm{Var}(ax+b) = a^2 \cdot \mathrm{Var}(x)$$
$$\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y)$$

$$\mathrm{Var}(m) = \mathrm{Var}\left(\frac{1}{N}\sum_{t=1}^{N} x^t\right) = \frac{1}{N^2}\sum_{t=1}^{N} \mathrm{Var}\left(x^t\right) = \frac{N \cdot \sigma^2}{N^2} = \frac{\sigma^2}{N} \xrightarrow{N=\infty} 0$$

Evaluating an Estimator (cont.)

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

$$s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2$$

$$\sum_{t=1}^{N} x^t = N \cdot m$$

$$\begin{aligned}
E\left[s^2\right] &= E\left[\frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2\right] \\
&= E\left[\frac{1}{N}\sum_{t=1}^{N}\left(\left(x^t\right)^2 - \underline{2x^t \cdot m} + m^2\right)\right] \\
&= E\left[\frac{\left(\sum_{t=1}^{N}\left(x^t\right)^2\right) - \underline{2N \cdot m^2} + Nm^2}{N}\right] \\
&= E\left[\frac{\left(\sum_{t=1}^{N}\left(x^t\right)^2\right) - N \cdot m^2}{N}\right] \\
&= \frac{\left(\sum_{t=1}^{N} E\left[\left(x^t\right)^2\right]\right) - N \cdot E\left[m^2\right]}{N}
\end{aligned}$$

Evaluating an Estimator (cont.)

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

$$\mathrm{Var}(m) = \frac{\sigma^2}{N} = E[m^2] - (E[m])^2$$
$$\Rightarrow E[m^2] = \frac{\sigma^2}{N} + (E[m])^2 = \frac{\sigma^2}{N} + \mu^2$$

$$E[s^2] = \frac{\left(\sum_{t=1}^{N} E\left[\left(x^t\right)^2\right]\right) - N \cdot E[m^2]}{N}$$

$$= \frac{N\left(\sigma^2 + \mu^2\right) - N\left(\frac{\sigma^2}{N} + \mu^2\right)}{N}$$

$$= \frac{(N-1)}{N}\sigma^2 \xrightarrow{N=\infty} \sigma^2$$

$$\mathrm{Var}(x^t) = \sigma^2 = E\left[\left(x^t\right)^2\right] - \left(E\left[x^t\right]\right)^2$$
$$\Rightarrow E\left[\left(x^t\right)^2\right] = \sigma^2 + \left(E\left[x^t\right]\right)^2 = \sigma^2 + \mu^2$$

The size of the observed sample set

Prior Information

- Given a function $g(X|\theta)$ (e.g., likelihood density $P(X|\theta)$) with parameter θ to be estimated
  - The prior density $P(\theta)$ tells some prior information on the possible value range that θ may take is helpful
    - Especially when the set of training samples is small
    - θ is treated as a random variable and $P(\theta)$ tells the likely values that θ may take before looking at the samples
    - The parameters in $P(\theta)$ are called hyperparameters
  - The prior density $P(\theta)$ can be combined with the likelihood density $P(X|\theta)$ to form the posterior density of θ

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta')P(\theta')d\theta'}$$

Prior Information

- Conjugate Priors
  - A prior density $P(\theta)$ which can make the posterior density $P(\theta|\mathbf{X})$, likelihood density $P(\mathbf{X}|\theta)$ and the prior density itself $P(\theta)$ belong to the same distribution family
    - The Gaussian (normal) density family

Prior Information (cont.)

- Example 2
  - θ is approximately normal $\Rightarrow$ $P(\theta) \sim N(\mu, \sigma^2)$
  - θ lies between 5 and 9, symmetrically around 7 with 90 percent confidence

$$P\{5 < \theta < 9\} = 0.9; \quad P(\theta) \sim N(\mu, \sigma^2), \ \mu = 7$$

$$\Rightarrow P\left\{-1.64 < \frac{\theta - \mu}{\sigma} < 1.64\right\} = 0.9 \ \ (\overline{N}(0, 1))$$

$$\Rightarrow P\{\mu - 1.64\sigma < \theta < \mu + 1.64\sigma\} = 0.9$$

$$\text{Take } \ 1.64\sigma = 2 \Rightarrow \sigma = 2/1.64$$

$$\Rightarrow P(\theta) \sim N\left(7, (2/1.64)^2\right)$$

Posterior Density

- The posterior density $P(\theta|\boldsymbol{X})$ of parameters θ tells the likely values after looking at the samples $\boldsymbol{X}$

$$P(\theta|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{P(\boldsymbol{X})} = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{\int P(\boldsymbol{X}|\theta')P(\theta')d\theta'}$$

Density/Output Estimation

- Density estimation of x using $P(x|\theta)$ and $P(\theta|\boldsymbol{X})$

$$P(x|\boldsymbol{X}) = \int P(x,\theta|\boldsymbol{X})d\theta$$

$$= \int \underline{P(x|\theta,\boldsymbol{X})}P(\theta|\boldsymbol{X})d\theta$$

$$= \int \underline{P(x|\theta)}P(\theta|\boldsymbol{X})d\theta$$

θ : **sufficient statistics**

- Output estimation of x using $g(x|\theta)$ and $P(\theta|\boldsymbol{X})$

$\hat{y} = g(x|\theta)$

$$y = \int g(x|\theta)P(\theta|\boldsymbol{X})d\theta$$

- Take an average over predictions ($P(x|\theta)$ or $g(x|\theta)$) using all value of θ , weighted by their (prior) probabilities

MAP and ML Estimators

- However, evaluating the integrals in above equations are not feasible
  - Estimation based a single point (point estimators)
    - Maximum A Posteriori Estimation
    - Maximum Likelihood Estimation
- Maximum A Posteriori (MAP) Estimator

$$P(x|\boldsymbol{X}) = P(x|\theta_{MAP}) \Rightarrow \quad \theta_{MAP} = \arg\max_{\theta} P(x|\theta)P(\theta|\boldsymbol{X})$$

$$g(x|\boldsymbol{X}) = g(x|\theta_{MAP}) \Rightarrow \quad \theta_{MAP} = \arg\max_{\theta} g(x|\theta)P(\theta|\boldsymbol{X})$$

- Maximum Likelihood (ML) Estimator

$$\theta_{ML} = \arg\max_{\theta} P(x|\theta)$$

$$\theta_{ML} = \arg\max_{\theta} g(x|\theta)$$

Bayes' Estimators

- Bayes' Estimator
  - Defined as the expected value of θ given its posterior density is known

$$\theta_{Bayes} = E[\theta|X] = \int \theta \cdot P(\theta|\mathbf{X})d\theta$$

  - Suppose that $E[\theta|X] = \mu$ and a estimator with value c is made
    - Mean square error of the estimator

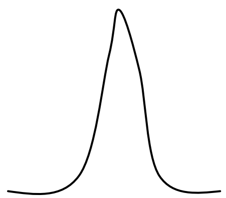
$$E\left[(\theta - c)^2|X\right] = E\left[((\theta - \mu) + (\mu - c))^2|X\right]$$
$$= E\left[(\theta - \mu)^2 + 2(\theta - \mu)(\mu - c) + \underline{(\mu - c)^2}|X\right]$$
$$= E\left[(\theta - \mu)^2|X\right] + (\mu - c)^2$$

constant

Bayes' Estimators (cont.)

- Bayes' Estimator (cont.)
  - Mean square error is minimum when $c = \mu = \theta_{Bayes}$
  - The best estimator of a random variable is its mean

- If the likelihood density $P(X|\theta)$ and the prior density $P(\theta)$ belong to normal densities
  - $P(\theta|X)$ is also normal
  - $\theta_{Bayes} \overset{?}{=} \theta_{MAP}$



$P(\theta|X)$

Bayes' Estimators (cont.)

- Example 2:
  - Given the likelihood density $P(\boldsymbol{X}|\theta)$ and the prior density $P(\theta)$ belong to normal densities

$$p(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(x-\theta)^2}{2\sigma_0^2}\right]$$

θ : variable

σ^2 : fixed constant

$$P(\boldsymbol{X}|\theta)=\frac{1}{(2\pi)^{N/2}\sigma_0^N}\exp\left[-\frac{\sum_t (x^t-\theta)^2}{2\sigma_0^2}\right]$$

$$P(\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right]$$

  - What is the estimate $\theta_{Bayes}=E[\theta|X]$?

$$E[\theta|\boldsymbol{X}]=\theta_{Bayes}=\frac{N/\sigma_0^2}{N/\sigma_0^2+1/\sigma^2}m+\frac{1/\sigma^2}{N/\sigma_0^2+1/\sigma^2}\mu$$

?

- m sample mean
- σ^2 sample variance

Parametric Classification

- Bayes' Classifier Revisited

x is assumed to be univariate

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)} = \frac{P(x|C_i)P(C_i)}{\sum_{k=1}^{K} P(x|C_k)P(C_k)}$$

  - Use discriminant function

$$g_i(x) = P(x|C_i)P(C_i)$$

denominator is dropped

$$\Rightarrow g_i(x) = \log P(x|C_i) + \log P(C_i)$$

logarithm is monotonic

  - How can we interpret $P(x|C_i)$ and $P(C_i)$?

Parametric Classification (cont.)

- Bayes' Classifier Revisited
  - Assume $P(x|C_i)$ is Gaussian

$$P(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

  - $P(C_i)$ is simply the proportion of samples x that belong to C_i

$$\Rightarrow g_i(x) = -\frac{1}{2}\log 2\pi - \log\sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Parametric Classification (cont.)

- Bayes' Classifier Revisited
  - How can we estimate $P(x|C_i)$ and $P(C_i)$?
    - Perform maximum like estimation (MLE) on the given (training) samples

$$X = \{x^t, \boldsymbol{r}^t\}_{t=1}^N, \quad r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_k, k \neq i \end{cases}$$

$$\hat{\mu}_i = m_i (\text{sample mean}) = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \qquad \hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

$$\hat{\sigma}_i^2 = \hat{s}_i^2 (\text{sample variance}) = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

$\left(\sum_t r_i^t\right) - 1$??

$$\Rightarrow g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2 s_i^2} + \log \hat{P}(C_i)$$

- How about Bayesian or MAP estimation ?

Parametric Classification (cont.)

- Bayes' Classifier Revisited
  - If class variances and priors are further set to be equal among the classes

$$\Rightarrow \hat{g}_i(x) = -(x - m_i)^2$$

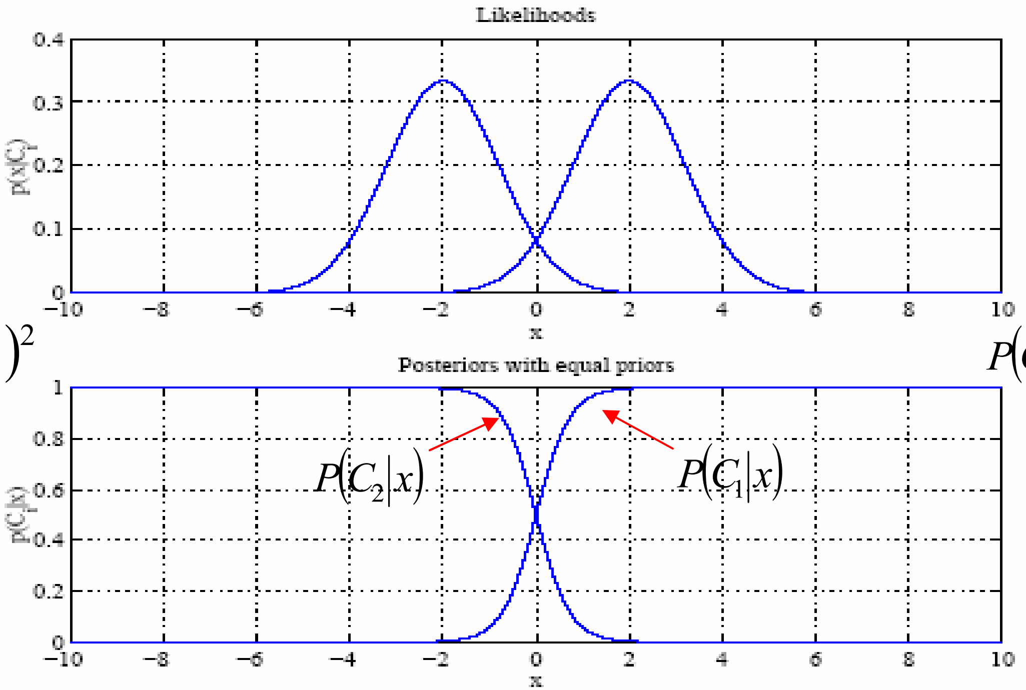
$$\text{Choose } C_\text{i} \text{ if } |x - m_i| = \min_k |x - m_k|$$

  - Assign x to the class with the nearest mean

Parametric Classification (cont.)

- E.g., Classes with Equal Priors and Variances

$$g_1(x) = g_2(x)$$
$$(x - m_1)^2 = (x - m_2)^2$$
$$x = \frac{m_1 + m_2}{2}$$



$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{P(x|C_1)}{P(x|C_1) + P(x|C_2)}$$

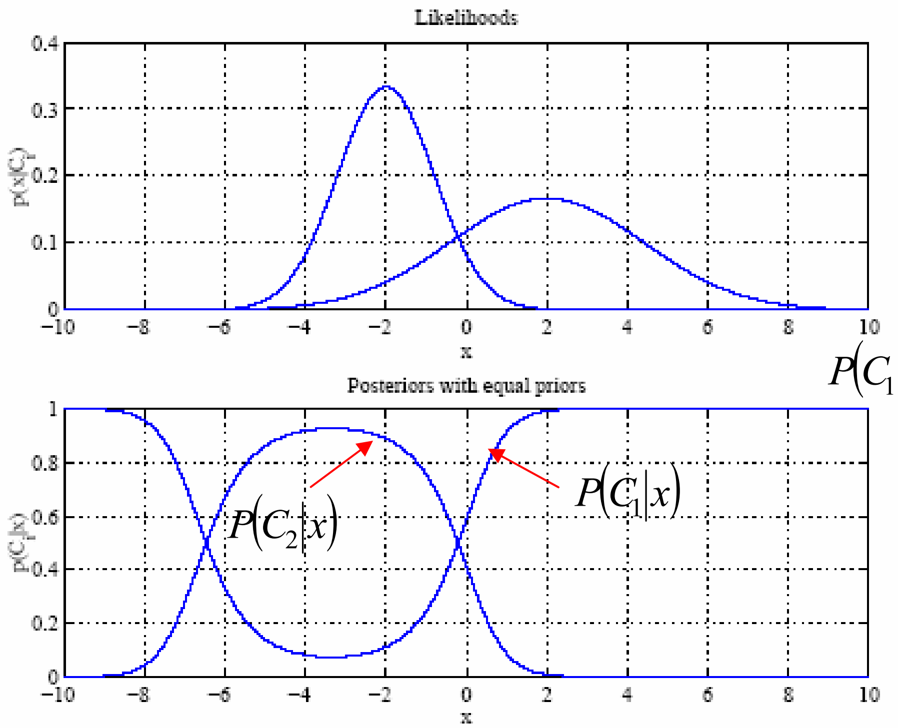
Figure 4.2: Likelihood functions and the posteriors with equal priors for two classes when the input is one-dimensional. Variances are equal and the posteriors intersect at one point, which is the threshold of decision.

Parametric Classification (cont.)

- E.g., Classes with Equal Priors and Unequal Variances

C_1 has a larger variance

?



$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{P(x|C_1)}{P(x|C_1) + P(x|C_2)}$$

Figure 4.3: Likelihood functions and the posteriors with equal priors for two classes when the input is one-dimensional. Variances are unequal and the posteriors intersect at two points.

Parametric Classification (cont.)

- Two common approaches for classification problems
  - Likelihood-based approach (as the classifiers mentioned above)
    - Estimate the probability distribution (likelihood density) for samples $P(x|C_i)$
    - Get the discriminant function using Bayes' rule $g_i(x) = P(x|C_i)P(C_i)$
    - Gaussian densities are usually assumed for continuous variables
      - Normality test is needed : bell-shaped (unimodal, symmetric around the center)
    - Example classifiers: HMMs
  - Discriminant-based approach
    - Bypass the estimation of densities and directly estimate the discriminants $e.g., \; g_i(x) = ax + b$
    - Example classifiers: Neural Networks, Support Vector Machines

Regression

- Function Approximation
  - Assume the observed numeric output is the sum of a deterministic function of the input and random noise

$$r = f(x) + \varepsilon$$

$$f \text{ is a fixed but unkown function; } \varepsilon \sim N(0, \sigma^2)$$

  - An estimator $g(x|\theta)$ used to approximate $f(x)$
    - θ : a set of parameters

$$r \sim N(g(x|\theta), \sigma^2)$$

$$\Rightarrow P(r|x) = N(r; g(x|\theta), \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{[r - g(x|\theta)]^2}{2\sigma^2}\right]$$

Regression (cont.)

- Find a parameter setting θ of $g(x|\theta)$ that can maximize the logarithm of the product of the likelihoods $P(r^t|x^t)$ for all training samples $X = \{x^t, \boldsymbol{r}^t\}_{t=1}^{N}$

$$L(\theta|X) = \log \prod_{t=1}^{N} P(r^t|x^t)$$

$$= \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{[r^t - g(x^t|\theta)]^2}{2\sigma^2}\right]$$

$$= \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\right] - \frac{1}{2\sigma^2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

constant

- Maximizing $L(\theta|X)$ is equivalent to minimizing the error function

$$E(\theta|X) = \frac{1}{2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

Least squares estimation
(minimizing the sum of squared errors)

Regression (cont.)

- Further assume that is a linear model (model)
  - Linear Regression

$$g\left(x^t \middle| \theta\right) = w_1 \cdot x^t + w_0$$

$$\Rightarrow E\left(\theta \middle| X\right) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - (w_1 \cdot x^t + w_0)\right]^2$$

- Find the minimum of $E\left(\theta \middle| X\right)$ by taking partial derivatives with respect to w_0 and w_1 accordingly

$$\frac{\partial E\left(\theta \middle| X\right)}{\partial w_0} = -\sum_{t=1}^{N} \left[r^t - (w_1 \cdot x^t + w_0)\right] = 0 \Rightarrow \sum_{t=1}^{N} r^t = N \cdot w_0 + w_1 \cdot \sum_{t=1}^{N} x^t$$

$$\frac{\partial E\left(\theta \middle| X\right)}{\partial w_1} = -\sum_{t=1}^{N} \left[r^t - (w_1 \cdot x^t + w_0)\right] \cdot x^t = 0 \Rightarrow \sum_{t=1}^{N} \left(r^t \cdot x^t\right) = w_0 \cdot \sum_{t=1}^{N} x^t + w_1 \cdot \sum_{t=1}^{N} \left(x^t\right)^2$$

Regression (cont.)

- Express the above two equations in vector-matrix form

$$\underbrace{\begin{bmatrix} N & \sum_{t=1}^{N} x^t \\ \sum_{t=1}^{N} x^t & \sum_{t=1}^{N} \left(x^t\right)^2 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\boldsymbol{w}} = \underbrace{\begin{bmatrix} \sum_{t=1}^{N} r^t \\ \sum_{t=1}^{N} \left(r^t \cdot x^t\right) \end{bmatrix}}_{\boldsymbol{y}}$$

$$\Rightarrow \boldsymbol{w} = \boldsymbol{A}^{-1}\boldsymbol{y} \quad \text{(if the inverse of } \boldsymbol{A} \text{ exists)}$$

- Also can be extended to polynomial regression

$$g\left(x^t \middle| \theta\right) = w_k \cdot \left(x^t\right)^k + \cdots + w_2 \cdot \left(x^t\right)^2 + w_1 \cdot \left(x^t\right) + w_0$$

HW-5: Parametric Classification

- Perform two-class classification (dichotomy) on the two data sets (MaleData, FemaleData) given in HW-3
  - The first 1000 samples coming from each data set are reserved for testing and are blended together
  - The rest samples of each data sets (training data sets) are respectively used to estimate the likelihood densities $P(x|C_i)$ and prior densities $P(C_i)$
  - All samples are projected onto the first eigenvector (dimension) of the LDA or PCA matrix
    - The LDA or PCA matrix are obtained merely based on the training data sets
      - Normality is assumed for the samples
  - Analyze the classification accuracy (using testing samples & Bayes's classifier) and also plot two figures regarding the likelihood densities $P(x|C_i)$ and poster densities $P(C_i|x)$ (using training samples)
    - As those shown in Fig. 4.3 (Alpaydin)