Parameter Estimation

Notational Convention

- Probabilities
  - Mass (discrete) function: capital letters
  - Density (continuous) function: small letters
- Vector vs. scalar
  - Scalar: plain
  - Vector: bold
  - 2D: small
  - Higher dimension: capital
- Notes in a continuous state of fluctuation until a topic is finished (many updates)

Parameter Estimation

- Optimal classifier maximizes
  - *a prior* probability
  - class-conditional density

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

- Assumption
  - no correlation
  - time independent statistics

Popular Approaches

- *Parametric*: assume a certain parametric form for $p(\mathbf{x}|w_i)$ and estimate the parameters
- *Nonparametric*: does not assume a parametric form for $p(\mathbf{x}|w_i)$ and estimate the density profile directly
- *Boundary*: estimate the separation hyperplane (hypersurface) between $p(\mathbf{x}|w_i)$ and $p(\mathbf{x}|w_j)$

a prior *probability*

- Given the numbers of occurrence:
  - if number of samples are large enough
  - the selection process is not biased
  - *Caveat*: sampling may be biased

$$(n_1, \varpi_1), (n_2, \varpi_2), \cdots, (n_k, \varpi_k)$$

$$\sum_{i=1}^{k} n_i = M$$

$$P(\varpi_i) = \frac{n_i}{M} \quad i = 1, \cdots, k$$

Class conditional density

- More complicated (not a single number, but a *distribution*)
  - assume a certain form
  - estimate the parameters
- What form should we assume?
  - Many, but in this course
  - We use almost exclusively Gaussian

Gaussian Distribution

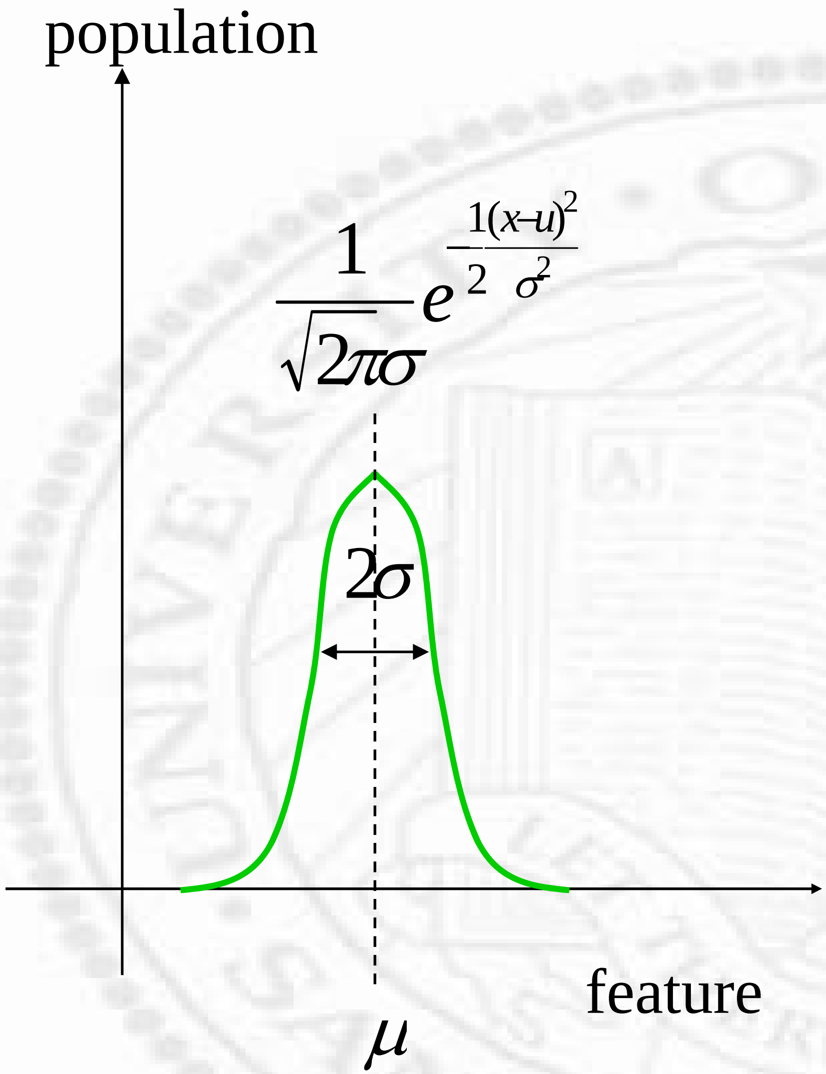
- Gaussian (or Normal) Scalar case

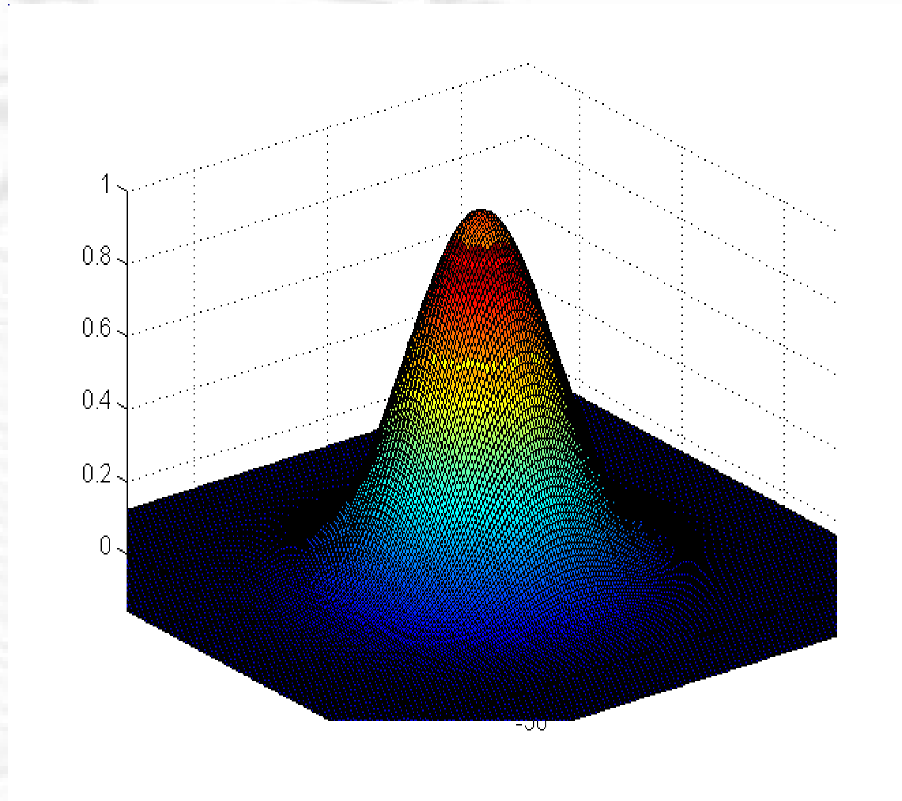
$$p(x\mid\varpi_i) = N(\mu_i,\sigma_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{1}{2}\frac{(x-u_i)^2}{\sigma_i^2}}$$

- Vector case

$$p(\mathbf{x}\mid\varpi_i) = N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = \frac{1}{\sqrt[d]{2\pi}\,|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}[(\vec{\mathbf{x}}-\vec{\mathbf{u}}_i)^T\boldsymbol{\Sigma}^{-1}(\vec{\mathbf{x}}-\vec{\mathbf{u}}_i)]}$$

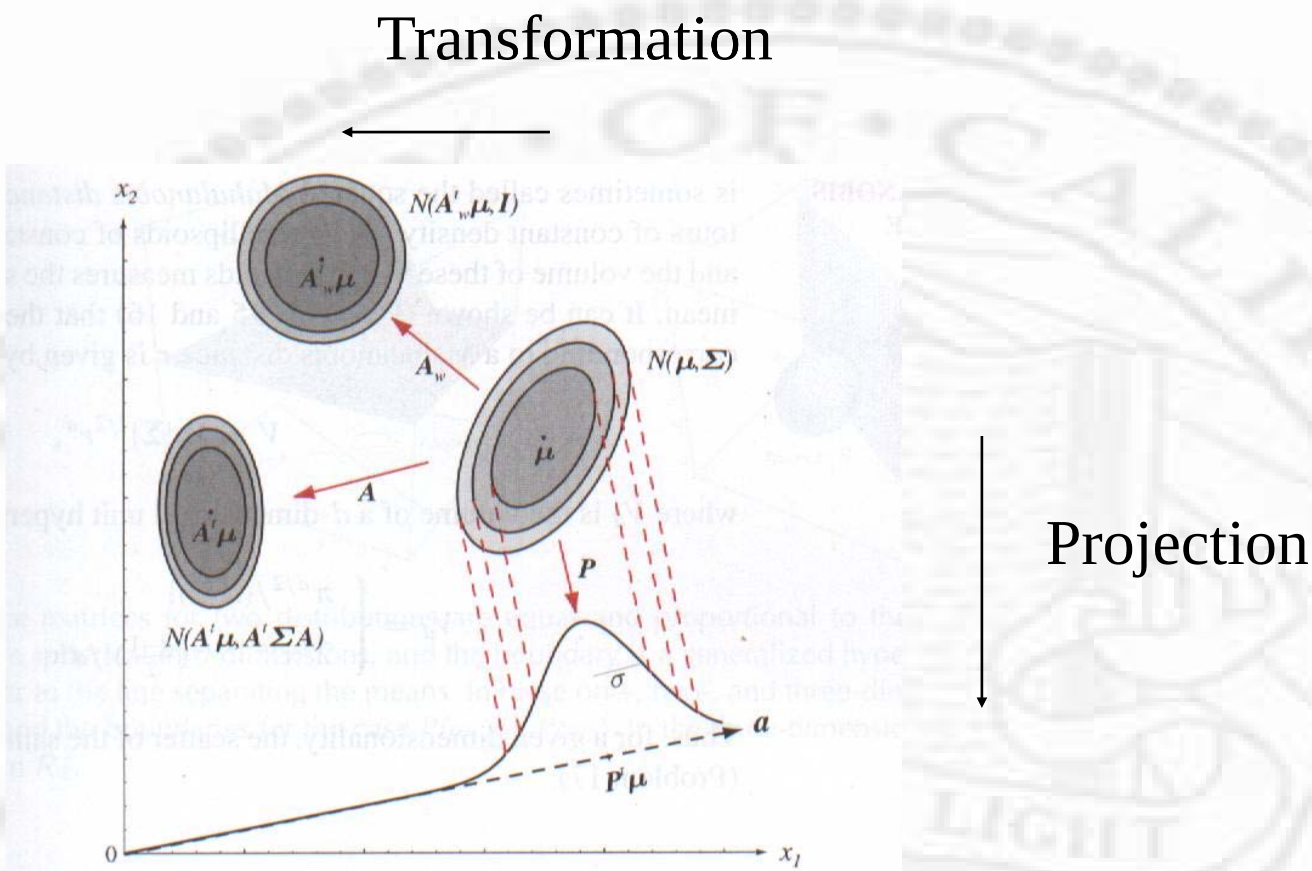
- Unknowns
  - class mean and variance


population
$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-u)^2}{\sigma^2}}$
2σ
μ
feature


1
0.8
0.6
0.4
0.2
0

Why Gaussian (Normal)?

- Central limit theorem predicts normal distribution from IID experiments
- In reality
  - There are only two numbers in the scalar case (mean and variance) to estimate, (or d + d(d+1)/2 in d-dimensions)
  - Nice mathematical properties (e.g., Fourier transform of a Gaussian is a Gaussian. Products and summation of Gaussian remain Gaussian, Any linear transform of a Gaussian is a Gaussian)



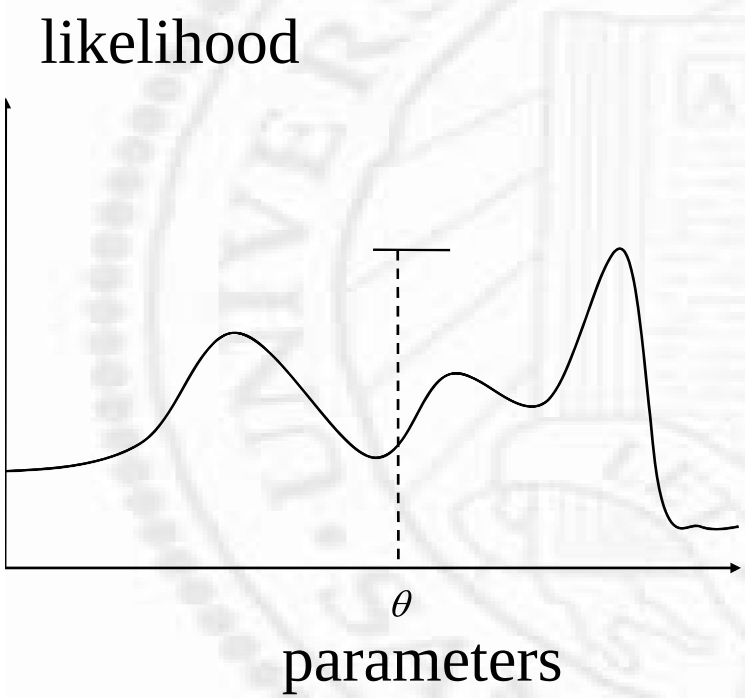
- In particular, a whitening transform can diagonalize the covariance matrix

Parameter Estimation

- Maximum likelihood estimator
  - Parameters have *fixed but unknown* values
- Bayesian estimator
  - parameters as *random variables* with know *a prior* distributions
  - Bayesian estimator allows us to change the a priori distribution by incorporating measurements to sharpen the profile

Graphically

- MLE
- Bayesian



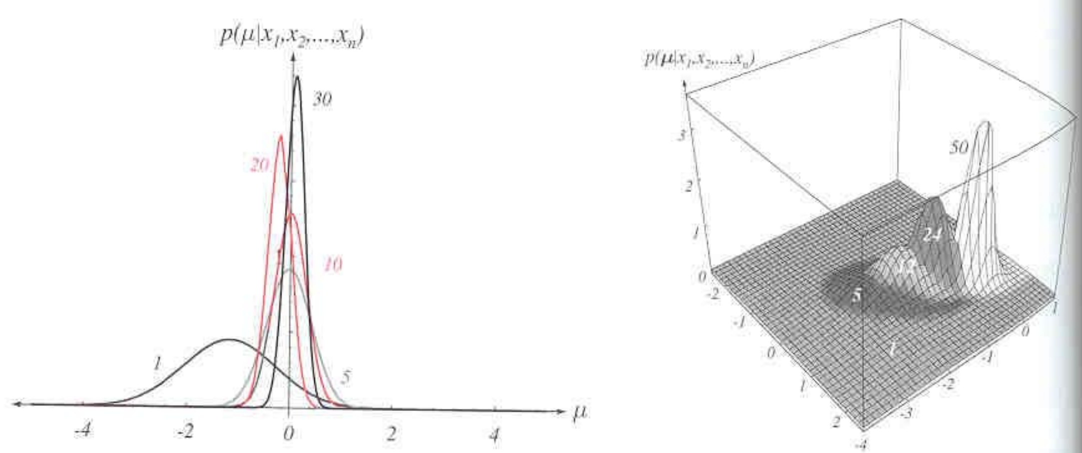


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation.

Maximum Likelihood Estimator

- Given
  - n labeled samples (observations)

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

  - an assumed distribution of *e* parameters

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_e\}$$

  - samples are drawn independently from

$$p(\mathbf{X}_j \mid \varpi) = p(\mathbf{X}_j \mid \boldsymbol{\theta}, \omega)$$

- Find
  - parameter that best explains the observations

MLE Formulation

- Maximize

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}_j|\boldsymbol{\theta})$$

Log likelihood

Or

$$l(\boldsymbol{\theta}) = \log p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \log p(\mathbf{x}_j|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_j|\boldsymbol{\theta}) = 0$$

An Example

$$p(x_j \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x_j - u)^2}{\sigma^2}}$$

$$\log p(x_j \mid \boldsymbol{\theta}) = -\frac{1}{2}\log \sigma^2 - \frac{1}{2}\frac{(x_j - u)^2}{\sigma^2}$$

$$\theta_1 = u \qquad \theta_2 = \sigma^2$$

$$\log p(x_j \mid \boldsymbol{\theta}) = -\frac{1}{2}\log \theta_2 - \frac{1}{2}\frac{(x_j - \theta_1)^2}{\theta_2}$$

$$\nabla_{\boldsymbol{\theta}} \log p(x_j \mid \boldsymbol{\theta}) = \begin{bmatrix} \dfrac{(x_j - \theta_1)}{\theta_2} \\ -\dfrac{1}{2\theta_2} + \dfrac{1}{2}\dfrac{(x_j - \theta_1)^2}{\theta_2^2} \end{bmatrix}$$

An Example (cont.)

$$\sum_{j=1}^{n} \frac{(x_j - \theta_1)}{\theta_2} = 0$$

$$\sum_{j=1}^{n} -\frac{1}{\theta_2} + \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{\theta_2^2} = 0$$

$$\hat{\mu} = \theta_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

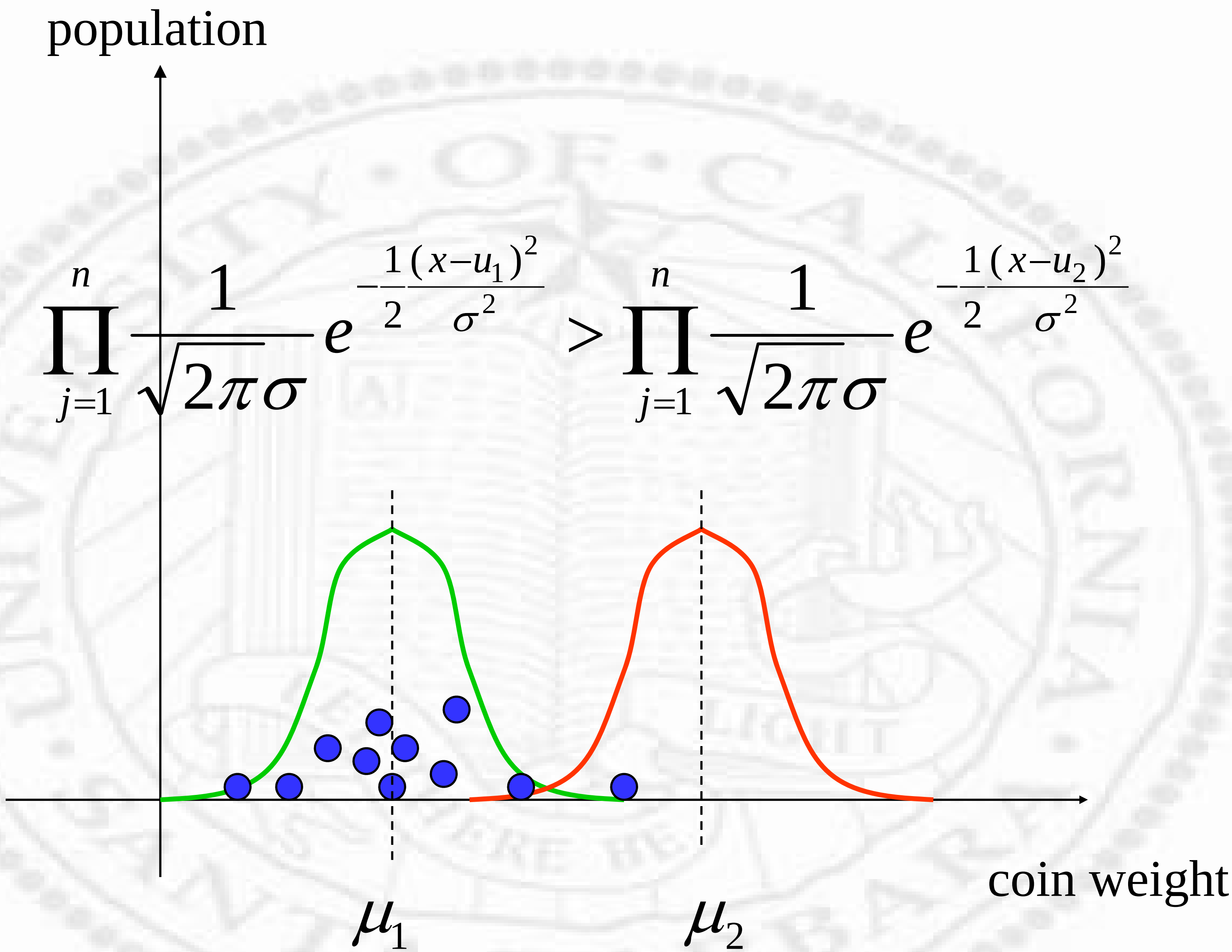
$$\hat{\sigma}^2 = \theta_2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \hat{u})^2$$

n-1, MLE is biased!

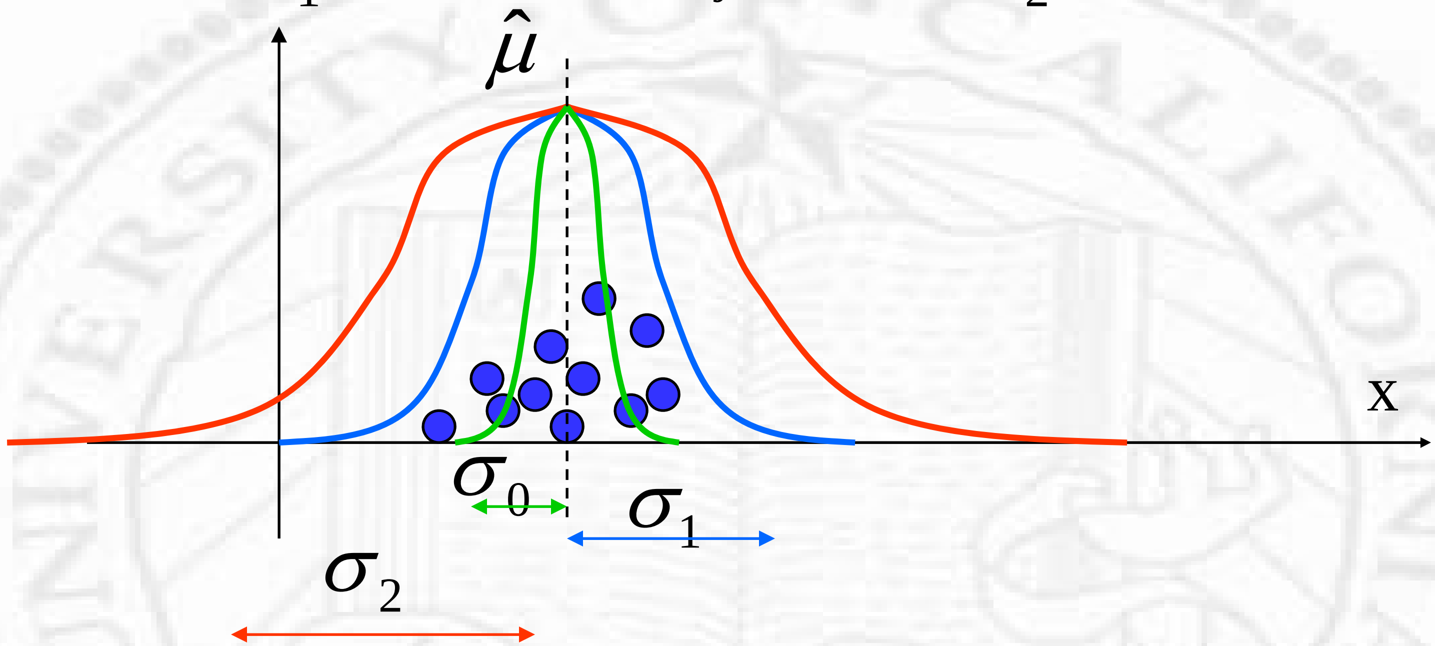
- Class mean as sample mean
- class variance as sample variance

$$g_i(x) = p(\varpi_i \mid x) = \frac{1}{p(x)} N(\hat{\mu}_i, \hat{\sigma}_i)\, p(\varpi_i) = \alpha \frac{1}{\sqrt{2\pi}\,\hat{\sigma}_i} e^{-\frac{1}{2}\frac{(x-\hat{u}_i)^2}{\hat{\sigma}_i^2}} p(\varpi_i)$$

$$\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-u_1)^2}{\sigma^2}} > \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-u_2)^2}{\sigma^2}}$$



$$\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(x-\hat{u})^2}{\sigma_1{}^2}} > \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\frac{(x-\hat{u})^2}{\sigma_2{}^2}}$$



- If σ too narrow, many sampling points will be outside 2 σ width with low likelihood of occurrence
- If σ too wide, 1/ σ becomes too small and reduces the likelihood of occurrence

A Quick Word on MAP

- MAP (Maximum *a posteriori*) estimator
- Similar to MLE with one additional twist
  - Maximize the (log) likelihood, l(.) *and*
  - p(.), prior probability of parameter values (if you know it), e.g., the mean is more likely to be u_o with a normal distribution
- MLE has a uniform prior, MAP not necessarily
- The added term is a case of "regularization"

Bayesian Estimator

- Note that MLE is a *batch* estimator
  - All data have to be kept
  - Difficult to update estimation
  - Difficult to incorporate other evidence
  - Insist on a single measurement
- Bayesian estimator
  - Allow the freedom that parameters in themselves can be random variables
  - Allow multiple evidence
  - Allow iterative update

Bayesian Estimator

- Based on Bayes rule

$$P(\varpi_i \mid x) = \frac{P(x, \varpi_i)}{P(x)} = \frac{p(x \mid \varpi_i) P(\varpi_i)}{\sum_j p(x \mid \varpi_j) P(\varpi_j)}$$

- With **X** at our disposal

$$P(\varpi_i \mid \mathbf{x}, \mathbf{X}) = \frac{P(\mathbf{x}, \varpi_i, \mathbf{X})}{P(\mathbf{x}, \mathbf{X})} = \frac{p(\mathbf{x} \mid \varpi_i, \mathbf{X}) P(\varpi_i \mid \mathbf{X})}{\sum_j p(\mathbf{x} \mid \varpi_j, \mathbf{X}) P(\varpi_j \mid \mathbf{X})}$$

Bayes Rule Formulation

- Assume
  - $\mathbf{X}$ comes from only one class
  - $p(\varpi_i)$ is independent of $\mathbf{X}$

$$p(\varpi_i \mid \mathbf{x}, \mathbf{X}) = \frac{P(\mathbf{x}, \varpi_i, \mathbf{X}_i)}{P(\mathbf{x}, \mathbf{X})} = \frac{p(\mathbf{x} \mid \varpi_i, \mathbf{X}_i) P(\varpi_i)}{\sum_j p(\mathbf{x} \mid \varpi_j, \mathbf{X}_j) P(\varpi_j)}$$

How can X be used?

- The distribution is *known* (e.g., normal), the parameters are *unknown*
- For estimating class parameters $p(\boldsymbol{\theta} \mid \mathbf{X})$
- class parameters then constrain x $p(\mathbf{x} \mid \boldsymbol{\theta})$
- put it all together

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

Bayes Rule Formulation (cont.)

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

- Ideally

$$p(\boldsymbol{\theta} \mid \mathbf{X}) = \begin{cases} 1 & some \, \hat{\boldsymbol{\theta}} \\ 0 & otherwise \end{cases}$$

This is MLE!

$$p(\mathbf{x} \mid \mathbf{X}) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \mid \mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$$

Otherwise, all possible θ's are used

Graphic Interpretation

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_e\}$$

$$\boxed{\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_e\} \quad \boldsymbol{\theta}' = \{\theta_1', \theta_2', \cdots, \theta_e'\} \quad \boldsymbol{\theta}'' = \{\theta_1'', \theta_2'', \cdots, \theta_e''\}}$$

$$p(\mathbf{x} \mid \boldsymbol{\theta}) P(\varpi)$$

An example

- Estimating mean of a normal distribution
- Variance is known
- Using n samples
- First step

Key to Bayesian: Both current and prior evidence can be used

$$p(x\,|\,\mathbf{X}) = \int p(x\,|\,u)\boxed{p(u\,|\,\mathbf{X})}du$$

$$\boxed{p(\mu\,|\,\mathbf{X})} = \frac{\boxed{p(\mathbf{X}\,|\,\mu)\,p(u)}}{p(\mathbf{X})}$$

Current evidence

$$p(\mathbf{X}\,|\,\mu) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}}$$

Previous and other evidence

$$p(u) = N(\mu_o, \sigma_o) = \frac{1}{\sqrt{2\pi}\,\sigma_o} e^{-\frac{1}{2}\frac{(\mu-\mu_o)^2}{\sigma_o^2}}$$

- Then

$$p(\mu \mid \mathbf{X}) = \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma_o} e^{-\frac{1}{2}\frac{(\mu-\mu_o)^2}{\sigma_o^2}}$$

$$= \alpha' e^{-\frac{1}{2}\{(\frac{n}{\sigma^2}+\frac{1}{\sigma_o^2})\mu^2 - 2(\frac{n}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2})\mu\}} = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_n^2}}$$

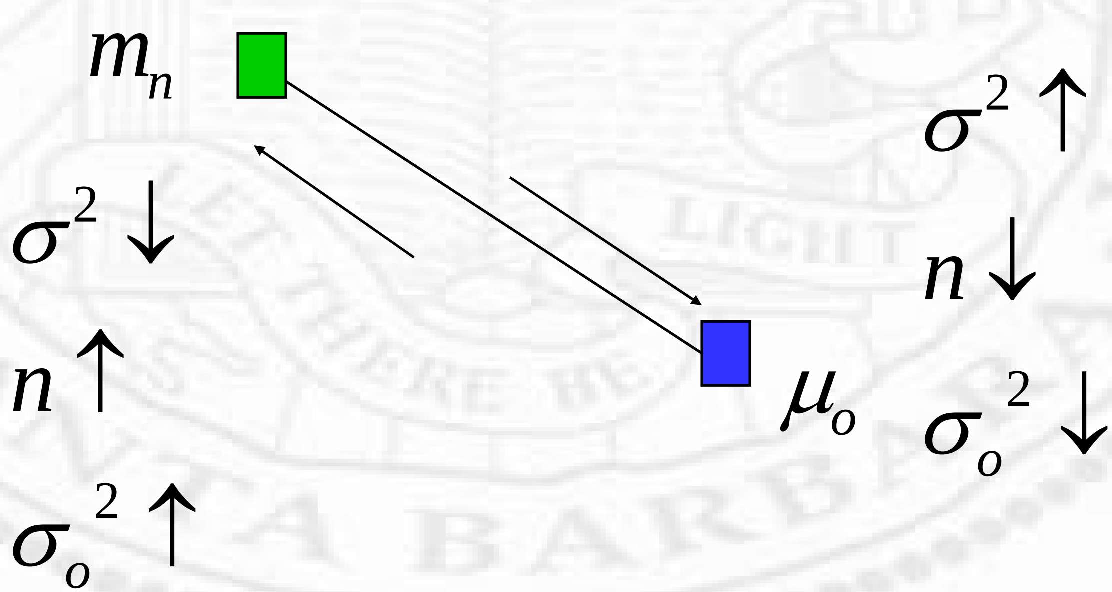
$$\mu_n = \frac{n\sigma_o^2}{n\sigma_o^2+\sigma^2} m_n + \frac{\sigma^2}{n\sigma_o^2+\sigma^2}\mu_o$$

$$\sigma_n^2 = \frac{\sigma_o^2\sigma^2}{n\sigma_o^2+\sigma^2} \qquad if \;\; \sigma^2 = \sigma_o^2 \Rightarrow \sigma_n^2 = \frac{\sigma^2}{n+1}$$

$$m_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

X helps in

- Defining the mean
- Reducing the uncertainty in mean
- Trust new data if
  - Class variance is small $\sigma^2 \downarrow$
  - Number of sample is large $n \uparrow$
  - Prior is uncertain $\sigma_o^2 \uparrow$



An example (cont.)

- Second step

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- Third step

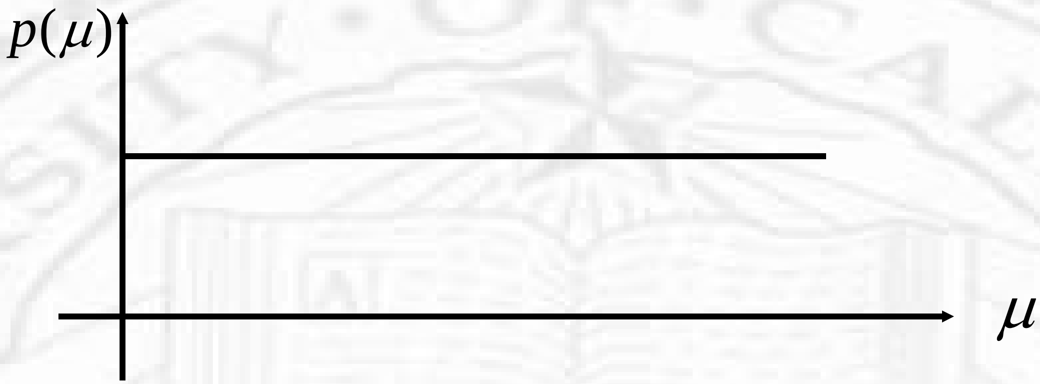
$$g(x) = p(x \mid \mathbf{X}) = \int p(x \mid \mu) p(u \mid \mathbf{X}) d\mu$$

$$= \int \{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\frac{(u-\mu_n)^2}{\sigma_n^2}} \} d\mu$$

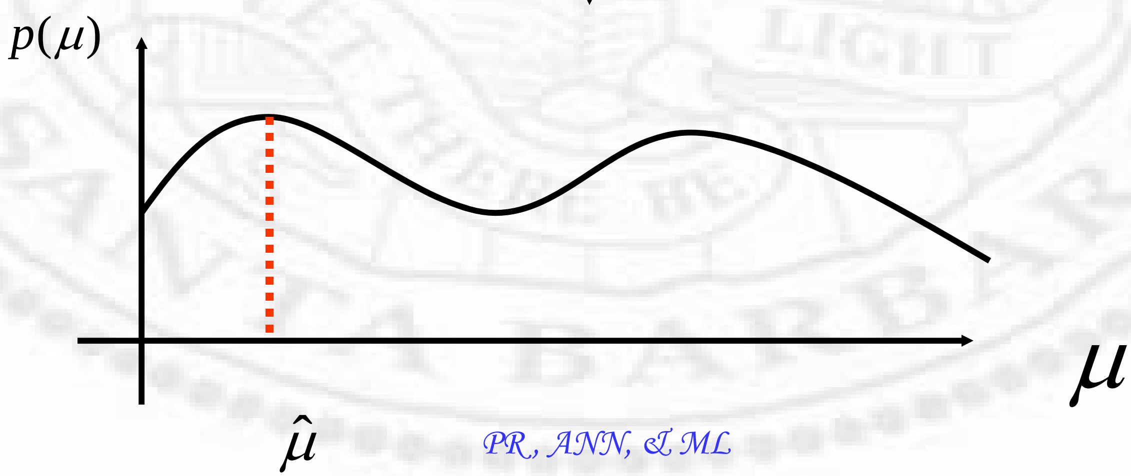
$$= N(\mu_n, \sigma^2 + \sigma_n^2) f(\sigma, \sigma_n)$$

$$\textit{where} \qquad f(\sigma, \sigma_n) = \int \exp\{ -\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} (u - \frac{\sigma^2 \mu_n + \sigma_n^2 x}{\sigma^2 + \sigma_n^2}) \} d\mu$$

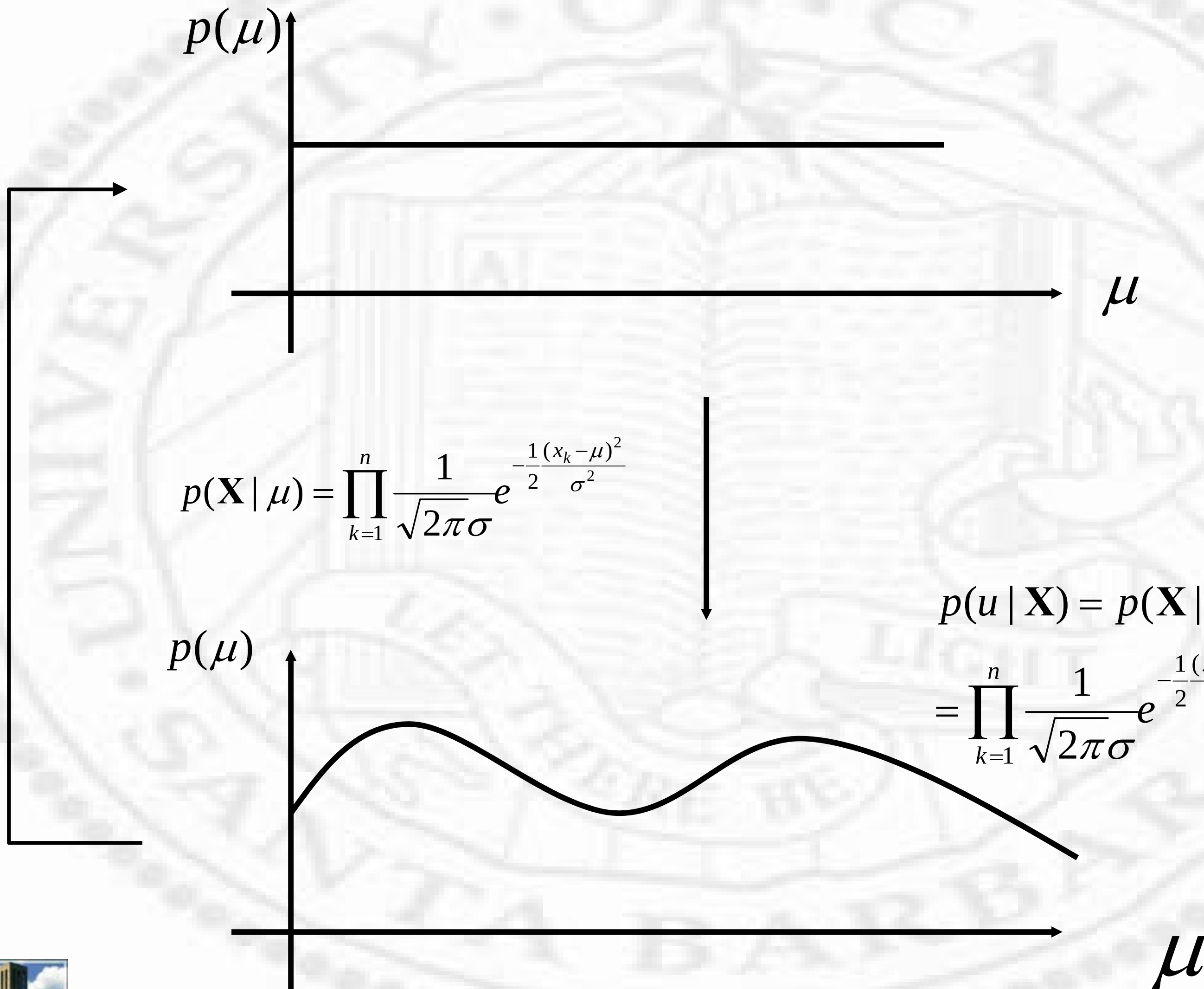
Graphical Interpretation: MLE



$$p(\mathbf{X} \mid \mu) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}}$$



Graphical Interpretation: Bayesian



$$p(\mathbf{X}\,|\,\mu)=\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}}$$

$$p(u\,|\,\mathbf{X})=p(\mathbf{X}\,|\,u)\,p(u)$$

$$=\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}}\;p(u)$$

Results of Iterative Process

- Start with a prior distribution
- Incorporate current batch of data
- Generate a new prior
- Goodness of new prior = goodness of old prior * goodness of interpretation
- Usually
  - Prior distribution sharpen (Bayesian learning)
  - Uncertainty drops

MLE vs. Bayes

- Faster (differentiation)
- Single model
- Known model $p(x|\theta)$
- Less information

- Slow (integration)
- Multiple weighted
- Unknown model fine
- More information (nonuniform prior)

Does it really make a difference?

- Yes, Bayesian classifier and MAP will in general give different results when used to classify *new* samples
- Because MAP (MLE) keeps only one hypothesis while Bayesian keeps multiple, weighted hypotheses

Example

- MLE

$$p(\mathbf{x}'|\mathbf{X}) = \arg\max_{\mathbf{x}} \; p(\mathbf{x}|\boldsymbol{\theta}'),$$

$$where \quad \boldsymbol{\theta}' = \arg\max_{\theta} \; P(\boldsymbol{\theta}|\mathbf{X})$$

- Bayesian

$$p(\mathbf{x}'|\mathbf{X}) = \arg\max_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathbf{X})\,d\boldsymbol{\theta}$$

$$p(\theta_1|\mathbf{X}) = .4, P(+|\theta_1) = 0, P(-|\theta_1) = 1$$

$$p(\theta_2|\mathbf{X}) = .3, P(+|\theta_2) = 1, P(-|\theta_2) = 0$$

$$p(\theta_3|\mathbf{X}) = .3, P(+|\theta_3) = 1, P(-|\theta_3) = 0$$

$$p(\mathbf{x}|\mathbf{X}) = -$$

Only one hypothesis (θ1) is kept

$$p(+|\mathbf{X}) = .4*0 + .3*1 + .3*1 = .6$$

$$p(-|\mathbf{X}) = .4*1 + .3*0 + .3*0 = .4$$

$$p(x|\mathbf{X}) = +$$

Gibbs Sampler

- Bayesian classifier is optimal, but can be very expensive – especially when a large number of hypotheses are kept and evaluated
- Gibbs – randomly pick one hypothesis according to the current posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X})$
- Can be shown (later) to be related knn classifier and the expected error is *at most twice* as bad as Bayesian

An Example: Naïve Bayesian

- Features are a conjunction of attributes
- Bayes theorem states that *a posterior* probability should be maximized
- Naïve Bayesian classifier assumes independence of attributes

$$
\begin{aligned}
c &= \arg\max_{c_j} P(c_j \mid a_1, a_2, \cdots, a_n) \\
&= \arg\max_{c_j} \frac{P(a_1, a_2, \cdots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \cdots, a_n)} \\
&= \arg\max_{c_j} P(c_j) \boxed{\prod_i P(a_i \mid c_j)}
\end{aligned}
$$

Example

Day	Outlook	Temperature	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Host	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cold	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example (cont)

- <Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>
- PlayTennis=yes? Or no?

$$c_{NB} = \arg\max_{c_j \in \{yes, no\}} P(c_j) P(Outlook = sunny \mid c_j) P(Temperature = cool \mid c_j)$$

$$P(Humidity = high \mid c_j) P(Wind = strong \mid c_j)$$

$$P(playTennis = yes) = \frac{9}{14} = .64$$

$$P(playTennis = no) = \frac{5}{14} = .36$$

$$P(Wind = strong \mid yes) = \frac{3}{9} = .33$$

$$P(Wind = strong \mid no) = \frac{3}{5} = .6$$

$$P(yes) P(sunny \mid yes) P(cool \mid yes)$$
$$P(high \mid yes) P(strong \mid yes) = 0.0053$$

$$P(no) P(sunny \mid no) P(cool \mid no)$$
$$P(high \mid no) P(strong \mid no) = 0.0206$$

Caveat

- Guarding against zero probability $P(a_i|c_j)$
  - Especially for small sample sizes and large set of attribute values
  - Use m-estimate instead
  - If attribute a_i can take k values, then $p=1/k$

$$p(a_i \mid c_j) = \frac{n_{a_i} + mp}{n_{c_j} + m}$$

n_{a_i} :# of samples in c_j with attribute a_i

n_{c_j} :# of samples in c_j

m : equivalent sample size (add m more samples)

p : prior estimate

More Examples

- Web page classification/Newsgroup classification
- Like/dislike for web pages
- Science/sports/entertainment categories for web pages/newsgroups

More Examples (cont.)

- Select common occurring words as features (at least *k* times in documents)
- Eliminate stop words (the, it, etc.) and punctuations
- Word stemming (like, liked etc.)
- $P(word_k | class_j)$ is independent of word position in the document
- Achieve 89% accuracy for classifying documents for 20 newsgroups