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

\[
p(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)}
\]

- Assumption
  - no correlation
  - time independent statistics
Popular Approaches

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

\[
\left(n_1, \omega_1\right), \left(n_2, \omega_2\right), \ldots, \left(n_k, \omega_k\right) \\
\sum_{i=1}^{k} n_i = M \\
P(\omega_i) = \frac{n_i}{M} \quad i = 1, \ldots, k
\]
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

\[ p(x | \omega_i) = N(\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2} \frac{(x-u_i)^2}{\sigma_i^2}} \]

- Vector case

\[ p(x | \omega_i) = N(\mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi|\Sigma_i|}} e^{-\frac{1}{2} [(x-\bar{u}_i)^T \Sigma_i^{-1} (x-\bar{u}_i)]} \]

- Unknowns

- class mean and variance
The equation for the population distribution is given by:

\[
\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
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 known *prior* distributions
  - Bayesian estimator allows us to change the *prior* distribution by incorporating measurements to sharpen the profile
Graphically

- MLE
- Bayesian

likelihood

parameters
Maximum Likelihood Estimator

- Given
  - n labeled samples (observations)
    \[ X = \{x_1, x_2, \ldots, x_n\} \]
  - an assumed distribution of \( e \) parameters
    \[ \theta = \{\theta_1, \theta_2, \ldots, \theta_e\} \]
  - samples are drawn independently from

- Find
  \[ p(X_j \mid \omega) = p(X_j \mid \theta, \omega) \]

- parameter that best explains the observations
MLE Formulation

Maximize

\[ p(X | \theta) = \prod_{j=1}^{n} p(x_j | \theta) \]

Log likelihood

\[ l(\theta) = \log p(X | \theta) = \sum_{j=1}^{n} \log p(x_j | \theta) \]

Or

\[ \nabla_{\theta} p(X | \theta) = 0 \]

\[ \nabla_{\theta} l(\theta) = \sum_{j=1}^{n} \nabla_{\theta} \log p(x_j | \theta) = 0 \]
An Example

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

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

\[ \theta_1 = u \quad \theta_2 = \sigma^2 \]

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

\[ \nabla_{\theta} \log p(x_j \mid \theta) = \begin{bmatrix} \frac{(x_j - \theta_1)}{\theta_2} \\ 2\theta_2 - \frac{1}{2} \frac{(x_j - \theta_1)^2}{\theta_2^2} \end{bmatrix} \]
An Example (cont.)

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

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

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

\[ \hat{\sigma}^2 = \theta_2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \hat{\mu})^2 \]

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

\[ g_i(x) = p(\omega_i \mid x) = \frac{1}{p(x)} N(\hat{\mu}_i, \hat{\sigma}_i) p(\omega_i) = \alpha \frac{1}{\sqrt{2\pi} \hat{\sigma}_i} e^{-\frac{1}{2} \left(\frac{(x-\hat{\mu}_i)^2}{\hat{\sigma}_i^2}\right)} p(\omega_i) \]
\[
\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-u_1)^2}{2\sigma^2}} > \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-u_2)^2}{2\sigma^2}}
\]
\[
\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2} \frac{(x-\hat{u})^2}{\sigma_1^2}} > \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_2}} e^{-\frac{1}{2} \frac{(x-\hat{u})^2}{\sigma_2^2}}
\]

- If \( \sigma \) too narrow, many sampling points will be outside 2 \( \sigma \) width with low likelihood of occurrence.
- If \( \sigma \) too wide, \( \frac{1}{\sigma} \) becomes too small and reduces the likelihood of occurrence.
A Quick Word on MAP

- MAP (Maximum \textit{a posteriori}) estimator
- Similar to MLE with one additional twist
  - Maximize the (log) likelihood, \( l(.) \) \textit{and}
  - \( p(.) \), prior probability of parameter values (if you know it), e.g., the mean is more likely to be \( \mu_0 \) 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
  \[ P(\omega_i \mid x) = \frac{P(x, \omega_i)}{P(x)} = \frac{p(x \mid \omega_i)P(\omega_i)}{\sum_j p(x \mid \omega_j)P(\omega_j)} \]

- With \( X \) at our disposal
  \[ P(\omega_i \mid x, X) = \frac{P(x, \omega_i, X)}{P(x, X)} = \frac{p(x \mid \omega_i, X)P(\omega_i \mid X)}{\sum_j p(x \mid \omega_j, X)P(\omega_j \mid X)} \]
Bayes Rule Formulation

- Assume
  
  - \( X \) comes from only one class
  - \( p(\omega_i) \) is independent of \( X \)

\[
p(\omega_i \mid x, X) = \frac{P(x, \omega_i, X_i)}{P(x, X)} = \frac{p(x \mid \omega_i, X_i)P(\omega_i)}{\sum_{j} p(x \mid \omega_j, X_j)P(\omega_j)}
\]
How can X be used?

- The distribution is known (e.g., normal), the parameters are unknown.
- For estimating class parameters \( p(\theta | X) \)
- class parameters then constrain \( x \) \( p(x | \theta) \)
- put it all together

\[
p(x | X) = \int p(x | \theta) p(\theta | X) d\theta
\]
Bayes Rule Formulation (cont.)

\[ p(x \mid X) = \int p(x \mid \theta)p(\theta \mid X)d\theta \]

- Ideally

\[
p(\theta \mid X) = \begin{cases} 
1 & \text{some } \hat{\theta} \\
0 & \text{otherwise} 
\end{cases} \quad \text{This is MLE!}
\]

\[ p(x \mid X) = \int p(x \mid \theta)p(\theta \mid X)d\theta = p(x \mid \hat{\theta}) \]

Otherwise, all possible \( \theta \)'s are used
Graphic Interpretation

\[ \theta = \{\theta_1, \theta_2, \ldots, \theta_e\} \]

\[ X = \{x_1, x_2, \ldots, x_n\} \]

\[ \theta' = \{\theta_1', \theta_2', \ldots, \theta_e'\} \]

\[ \theta'' = \{\theta_1'', \theta_2'', \ldots, \theta_e''\} \]

\[ p(x | \theta)P(\varnothing) \]
An example

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

\[ p(x \mid X) = \int p(x \mid u)p(u \mid X)du \]

\[ p(\mu \mid X) = \frac{p(X \mid \mu)p(u)}{p(X)} \]

Current evidence

\[ p(X \mid \mu) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x_k - \mu}{\sigma}\right)^2} \]

Previous and other evidence

\[ p(u) = N(\mu_o, \sigma_o) = \frac{1}{\sqrt{2\pi}\sigma_o} e^{-\frac{1}{2} \left(\frac{\mu - \mu_o}{\sigma_o}\right)^2} \]
Then

\[ p(\mu | X) = \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(x_k - \mu)^2}{\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma_o} e^{-\frac{1}{2} \frac{(\mu - \mu_o)^2}{\sigma_o^2}} = \alpha' e^{-\frac{1}{2} \left\{ \left( \frac{n}{\sigma^2} + \frac{1}{\sigma_o^2} \right) \mu^2 - 2 \left( \frac{n}{\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2} \right) \mu \right\}} \]

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

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

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  \( \sigma_o^2 \uparrow \)

\[ m_n \]
\[ \sigma^2 \downarrow \]
\[ n \uparrow \]
\[ \sigma^2 \uparrow \]
\[ \sigma_o^2 \uparrow \]

\[ \sigma^2 \uparrow \]
\[ n \downarrow \]
\[ \sigma_o^2 \downarrow \]
\[ \mu_o \]
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**

\[ g(x) = p(x \mid X) = \int p(x \mid \mu) p(u \mid X) \, d\mu \]

\[ = \int \left\{ \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} \cdot \frac{1}{\sqrt{2\pi \sigma_n}} e^{-\frac{1}{2} \frac{(u-\mu_n)^2}{\sigma_n^2}} \right\} \, d\mu \]

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

where

\[ f(\sigma, \sigma_n) = \int \exp\left\{ -\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} (u - \frac{\sigma^2 \mu_n + \sigma_n^2 x}{\sigma^2 + \sigma_n^2}) \right\} \, d\mu \]
Graphical Interpretation: MLE

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

\[ p(\mu) \]

\[ \hat{\mu} \]
Graphical Interpretation: Bayesian

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

\[ p(u | X) = p(X | u) p(u) \]

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

\[ p(\mathbf{x}' | \mathbf{X}) = \arg \max_{\mathbf{x}} p(\mathbf{x} | \theta') , \]

where \( \theta' = \arg \max_{\theta} P(\theta | \mathbf{X}) \)

\[ p(\theta_1 | \mathbf{X}) = .4, P(\cdot | \theta_1) = 0, P(- | \theta_1) = 1 \]

\[ p(\theta_2 | \mathbf{X}) = .3, P(\cdot | \theta_2) = 1, P(- | \theta_2) = 0 \]

\[ p(\theta_3 | \mathbf{X}) = .3, P(\cdot | \theta_3) = 1, P(- | \theta_3) = 0 \]

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

Only one hypothesis (\( \theta_1 \)) is kept

- Bayesian

\[ p(\mathbf{x}' | \mathbf{X}) = \arg \max_{\mathbf{x}} \int p(\mathbf{x} | \theta) p(\theta | \mathbf{X}) d\theta \]

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

\[ p(\cdot | \mathbf{X}) = .4 \times 0 + .3 \times 1 + .3 \times 1 = .6 \]

\[ p(- | \mathbf{X}) = .4 \times 1 + .3 \times 0 + .3 \times 0 = .4 \]
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 $p(\theta | x)$.
- 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 \textit{a posteriori} probability should be maximized
- Naïve Bayesian classifier assumes independence of attributes

\[
c = \arg \max_{c_j} P(c_j \mid a_1, a_2, \ldots, a_n) \\
= \arg \max_{c_j} \frac{P(a_1, a_2, \ldots, a_n \mid c_j)P(c_j)}{P(a_1, a_2, \ldots, a_n)} \\
= \arg \max_{c_j} P(c_j) \prod_i P(a_i \mid c_j)
\]
**Example**

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Play tennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cold</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Example (cont)

- <Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>

- PlayTennis=yes? Or no?

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

\[ P(\text{Humidity} = \text{high} \mid c_j)P(\text{Wind} = \text{strong} \mid c_j) \]

\[ P(\text{playTennis} = \text{yes}) = \frac{9}{14} = .64 \]

\[ P(\text{playTennis} = \text{no}) = \frac{5}{14} = .36 \]

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

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

\[ P(\text{yes})P(\text{sunny} \mid \text{yes})P(\text{cool} \mid \text{yes}) \]

\[ P(\text{high} \mid \text{yes})P(\text{strong} \mid \text{yes}) = 0.0053 \]

\[ P(\text{no})P(\text{sunny} \mid \text{no})P(\text{cool} \mid \text{no}) \]

\[ P(\text{high} \mid \text{no})P(\text{strong} \mid \text{no}) = 0.0206 \]
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.)
- $P(\text{word}_k | \text{class}_j)$ is independent of word position in the document
- Achieve 89% accuracy for classifying documents for 20 newsgroups