To obtain - Bph genus + ML estimators,
(Estimates of the other conditional prob.

Posterior and/or likelihood are not easy

is available (supervised learning is possible),

Even when knowledge of \( p(y|x) \) (aka. class labels)

dimensional data \( (x) \).

Small compared to a data set \( \not\subset \) large

In fact, what we actually really know is a

more known.

For, we assumed \( p(y|x) \) \( \not\subset \) \( \mathcal{L}(x) \)

CHAPTER 3

LECTURE 10
In the parametric case, if \( \Theta \) is given, the samples \( \{ X_i \} \) are independent and identically distributed (i.i.d.) from \( p(x|\Theta) \).

\[
\begin{align*}
(1) & \quad p(x|\Theta) \text{ is parametrized by } \Theta \\
(2) & \quad \{ X_i \} \text{ are i.i.d.}
\end{align*}
\]

The samples \( \{ X_i \} \) are drawn according to \( p(x|\Theta) \) with a different \( \Theta \) for each sample.

- \( \Theta \) does not change.

ML Estimators
\[ p(\Theta | \mathcal{D}) = \prod_{i=1}^{k} \frac{1}{k} \]

Also, given these in samples \( \mathcal{X} \), \( \mathcal{Y} \) in \( \mathcal{D} \), find \( \Theta \)

Given the samples \( \mathcal{X} \) in a dataset \( \mathcal{D} \), find \( \Theta \)

Neglect; we can drop the superscripts, and the problem reduces to:

As a separate problem:

Test each estimation

Recognizing \( \Theta \) (\( i \neq j \))

2 samples in \( \mathcal{D} \); give no information

Supervised learning

Test we know the labels

Of the datasets it's already successful

Use simple knowledge of the features

Assumptions:
Fixing $\theta_0$ & $\theta_1$ \(\forall n\)

Learning $\theta$ \(\forall n\)

Recall: \(P(\theta \mid x)\) is the likelihood of $\theta$ w.r.t. $x$.

The observations (samples) that are closest to $E$ are the best acceptable hypotheses' supports.

That is, $\theta$ corresponds to true value of $\theta$.

That vector $\theta$ that maximizes $P(x \mid \theta)$ is the ML estimate of $\theta$.

The ML estimation of $\theta$ is the likelihood of $\theta$ w.r.t. $x$. Which one? Best supports?
\[ \Theta = \arg \max_e \phi(\Theta) \]

Then

\[ \phi(\Theta) \leq \phi_e(\Theta) \]

\[
\begin{bmatrix}
\vdots \\
\frac{\Theta_{a1}}{a1} \\
\vdots \\
\frac{\Theta_{a0}}{a0}
\end{bmatrix} = \Delta
\]

So, let

**Analytically**

Examine and assess your current results.

- **Monotonicity**

Function of \(\Theta\) can be obtained.

If \(\Theta\) is a well-behaved, it is

- May or may not be well-behaved

*Notes:* \( p(\Theta) \) may or may not be well-behaved.
(\text{It is } \textit{good} \ \text{only when } n \to \infty)

\text{is an estimator of } \theta

\text{Inflection Points}

\text{Local/Global Maximum/Minimum}

\delta \Delta = 0

\text{is one of the solutions of}

\sum_{k=1}^{\infty} \theta_k \left( \frac{x_i}{\theta_k} \right) = 0

and

\max_{x_i} \left( \frac{x_i}{\theta_k} \right) = \max_{x_i} e^{x_i} \frac{\theta_k}{\theta} = \Theta
\[ x \leq \frac{1}{2} \quad \Rightarrow \quad 0 = ( - x ) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \Delta \]

\[ \left( \frac{1}{2} - x \right) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \Delta \]

\[ \left( \frac{1}{2} - x \right) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n = \left( \frac{1}{2} - x \right) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \]

\[ \sqrt{n} \quad \text{unknown} \quad (\frac{1}{2} - x) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \]

\[ \frac{1}{2} - x \quad \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \]

\[ \left( \frac{1}{2} - x \right) \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \]

\[ \frac{1}{2} - x \quad \frac{n}{\sqrt{2 \pi}} \left( \frac{1}{2} \right)^n \]

Normal distribution with \( \mu = \frac{1}{2} \).

Samples are drawn from a multivariate Gaussian distribution.
\[
\begin{bmatrix}
\frac{\varepsilon^2}{(\varepsilon - \theta)^2} & \varepsilon \\
\frac{\varepsilon}{(\varepsilon - \theta)^2} & 1
\end{bmatrix}
= \begin{bmatrix}
\varepsilon & 1 \\
1 & \frac{1}{\varepsilon}
\end{bmatrix}
= (\theta)^T \theta
\]

\[
\begin{bmatrix}
\frac{\varepsilon^2}{(\varepsilon - \theta)^2} \\
\frac{\varepsilon}{(\varepsilon - \theta)^2}
\end{bmatrix}
= \begin{bmatrix}
\varepsilon \\
1
\end{bmatrix}
= (\theta)^T \theta
\]

\[
\varepsilon (\theta - \theta')^2
\]

\[
\frac{\varepsilon^2}{(\varepsilon - \theta)^2} - \frac{\varepsilon}{(\varepsilon - \theta)^2} = \frac{z}{\varepsilon}
\]

\[
\begin{bmatrix}
\varepsilon^2 \\
\varepsilon
\end{bmatrix}
= \begin{bmatrix}
\varepsilon^2 \\
\varepsilon
\end{bmatrix}
= \theta
\]

Let's consider the univariate case:

\[
\exists \theta \in \mathbb{R}, \theta = \left[ \begin{array}{c}
\theta_1 \\
\theta_2
\end{array} \right]
\]

(2) unknown

\[
\{ \theta \}
\]
The matrix

\[ = \frac{1}{\sqrt{\Sigma}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) (x_i - \mu)^T \]

Note: \[ \Sigma = \frac{1}{T} = \frac{1}{n} \]

\[ \Rightarrow \left( \frac{1}{\sqrt{T}} \sum_{i=1}^{T} (x_i - \mu) (x_i - \mu)^T \right)^{1/2} \]

\[ \Rightarrow \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) (x_i - \mu)^T \right)^{1/2} \]

\[ \Rightarrow \frac{1}{\sqrt{T}} \sum_{i=1}^{T} (x_i - \mu) (x_i - \mu)^T \]

Back to the multivariate case (and after a few

\[ \frac{1}{\sqrt{T}} \sum_{i=1}^{T} (x_i - \mu) (x_i - \mu)^T = \Theta \]

\[ \Rightarrow \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) (x_i - \mu)^T = \Theta \]

\[ = 0 \]
\[ \sum_{i=1}^{n} \frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} \leq a \leq \sum_{i=1}^{n} \frac{1}{2} \]

\[ \sum_{i=1}^{n} \frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} \leq a \leq \sum_{i=1}^{n} \frac{1}{2} \]

Now,

\[ \frac{1}{\sum_{i=1}^{n} \frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} } \leq \frac{1}{a} = \frac{1}{2} \]

AND

\[ \frac{1}{\sum_{i=1}^{n} \frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} } \leq \frac{1}{a} = \frac{1}{2} \]

THEN

\[ \frac{1}{\sum_{i=1}^{n} \frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} } \leq \frac{1}{a} = \frac{1}{2} \]

Since

\[ \text{B isi (w.r.t. the real \( a \))} \]
e.g., we assume:

are not true?

What if other assumptions about \( p(x|\theta) \)

In order to eliminate bias, we use:

\[
\mathbb{E}[\theta] = \frac{1}{n-1} \sum_{i=1}^{n-1} (\hat{\theta}_i - \mu)^2
\]

Conductours pdf: \( p(x|\theta) \) the underlying
The co-matrix of the underlying
For Random partitions of the \( n \times m \) C

\[ n \text{-fold CV: Repeat the process above} \]

Test w/ remainder 10% and train the classifier. Cross validation: True, say 90% of the data.

\[ \text{Cross validation: True, say 90\% of the data} \]

Not the real world! Systems model the data.

Cons: Biased result. Because

Pros: Lots of training data

Use all data for training.

Resubstitution: Use all data for training & testing w/ real data
**Classification:**

$$ C_i \equiv \frac{N}{N} = 90\% \text{ Correct} $$

**Confusion Matrix:**

```
<table>
<thead>
<tr>
<th></th>
<th>C11</th>
<th>C12</th>
<th>C1m</th>
</tr>
</thead>
<tbody>
<tr>
<td>C21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2m</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cnm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Error Matrix:**

- **True Negatives:** $TN$
- **False Negatives:** $FN$
- **False Positives:** $FP$
- **True Positives:** $TP$

**Accuracy:**

$$ \frac{TP + TN}{TP + TN + FN + FP} $$

**Notations:**
- $TP$: True Positive
- $TN$: True Negative
- $FP$: False Positive
- $FN$: False Negative
Recall = \frac{TP}{TP + FN}

Precision = \frac{TP}{TP + FP}

True Positive Rate

Positive Predictive Value
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP}{TP + TN + FP + FN}
\]
\[
\text{False Positive Rate} (\text{FPR}) = \frac{FP}{FP + TN}
\]
\[
\text{False Negative Rate} (\text{FNR}) = \frac{FN}{FN + TP}
\]
Receiver Operating Characteristics

\[ \text{ROC} \]

\[ \text{FPR} = \frac{FP}{FP + TN} \]

\[ \text{TPR} = \frac{TP}{TP + FN} \]