Probability Basics, Density Estimation

These slides were assembled by Byron Boots, with only minor modifications from Eric Eaton’s slides and grateful acknowledgement to the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.
The Joint Distribution

Recipe for making a joint distribution of \( d \) variables:

1. Make a probability table listing all combinations of values of your variables (if there are \( d \) Boolean variables then the table will have \( 2^d \) rows).

2. For each combination of values, say how probable it is.

2. If you subscribe to the axioms of probability, those numbers must sum to 1.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

e.g., Boolean variables A, B, C

Slide © Andrew Moore
Inferring Marginal Probabilities from the Joint

<table>
<thead>
<tr>
<th></th>
<th>alarm</th>
<th>¬alarm</th>
<th>earthquake</th>
<th>¬earthquake</th>
<th>earthquake</th>
<th>¬earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>burglary</td>
<td>0.01</td>
<td>0.08</td>
<td>0.001</td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>¬burglary</td>
<td>0.01</td>
<td>0.09</td>
<td>0.01</td>
<td>0.79</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
P(\text{alarm}) = \sum_{b,e} P(\text{alarm} \land \text{Burglary} = b \land \text{Earthquake} = e)
\]

\[
= 0.01 + 0.08 + 0.01 + 0.09 = 0.19
\]

\[
P(\text{burglary}) = \sum_{a,e} P(\text{Alarm} = a \land \text{burglary} \land \text{Earthquake} = e)
\]

\[
= 0.01 + 0.08 + 0.001 + 0.009 = 0.1
\]
Conditional Probability

• $P(A \mid B) = \text{Probability that } A \text{ is true given } B \text{ is true}$

What if we already know that $B$ is true?

That knowledge changes the probability of $A$

• Because we know we’re in a world where $B$ is true

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

\[
P(A \land B) = P(A \mid B) \times P(B)
\]
Example: Conditional Probabilities

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

\[
P(A \land B) = P(A \mid B) \times P(B)
\]

\[
P(\text{Alarm, Burglary}) = \begin{array}{c|cc}
 & \text{alarm} & \neg\text{alarm} \\
\hline
\text{burglary} & 0.09 & 0.01 \\
\neg\text{burglary} & 0.1 & 0.8 \\
\end{array}
\]

\[
P(\text{burglary} \mid \text{alarm}) = \frac{P(\text{burglary} \land \text{alarm})}{P(\text{alarm})} = \frac{0.09}{0.19} = 0.47
\]

\[
P(\text{alarm} \mid \text{burglary}) = \frac{P(\text{burglary} \land \text{alarm})}{P(\text{burglary})} = \frac{0.09}{0.1} = 0.9
\]

\[
P(\text{burglary} \land \text{alarm}) = P(\text{burglary} \mid \text{alarm}) \cdot P(\text{alarm}) = 0.47 \times 0.19 = 0.09
\]
Example: Inference from Conditional Probability

\[ P(A \mid B) = \frac{P(A \land B)}{P(B)} \]
\[ P(A \land B) = P(A \mid B) \times P(B) \]

P(headache) = 1/10
P(flu) = 1/40
P(headache \mid flu) = 1/2

“Headaches are rare and flu is rarer, but if you’re coming down with the flu, then there’s a 50-50 chance you’ll have a headache.”

Based on slide by Andrew Moore
Example: Inference from Conditional Probability

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

\[
P(A \land B) = P(A \mid B) \times P(B)
\]

P(headache) = 1/10
P(flu) = 1/40
P(headache \mid flu) = 1/2

One day you wake up with a headache. You think: “Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu.”

Is this reasoning good?

Based on slide by Andrew Moore
Example: Inference from Conditional Probability

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

\[
P(A \land B) = P(A \mid B) \times P(B)
\]

P(headache) = 1/10
P(flu) = 1/40
P(headache \mid flu) = 1/2

Want to solve for:

P(headache \land flu) = ?
P(flu \mid headache) = ?

\[
P(headache \land flu) = P(headache \mid flu) \times P(flu)
\]

\[
= 1/2 \times 1/40 = 0.0125
\]

\[
P(flu \mid headache) = \frac{P(headache \land flu)}{P(headache)}
\]

\[
= 0.0125 / 0.1 = 0.125
\]

Based on example by Andrew Moore
Bayes’ Rule

\[
P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}
\]

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

\[
P(A \land B) = P(A \mid B) \times P(B)
\]
\[
P(B \land A) = P(B \mid A) \times P(A)
\]
these are the same

Just set equal...

\[
P(A \mid B) \times P(B) = P(B \mid A) \times P(A)
\]
and solve...
Bayes’ Rule

- Allows us to reason from evidence to hypotheses
- Another way of thinking about Bayes’ rule:

\[
P(\text{hypothesis} | \text{evidence}) = \frac{P(\text{evidence} | \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{evidence})}
\]

In the flu example:
- \( P(\text{headache}) = 1/10 \)
- \( P(\text{flu}) = 1/40 \)
- \( P(\text{headache} | \text{flu}) = 1/2 \)

Given evidence of headache, what is \( P(\text{flu} | \text{headache}) \)?

Solve via Bayes rule!
Independence

• When two sets of propositions do not affect each others’ probabilities, we call them **independent**

• Formal definition:

  \[
  A \perp B \iff P(A \land B) = P(A) \times P(B)
  \]

  \[
  \iff P(A \mid B) = P(A)
  \]

For example, \{moon-phase, light-level\} might be independent of \{burglary, alarm, earthquake\}

• Then again, maybe not: Burglars might be more likely to burglarize houses when there’s a new moon (and hence little light)

• But if we know the light level, the moon phase doesn’t affect whether we are burglarized
Exercise: Independence

<table>
<thead>
<tr>
<th>P(smart &amp; study &amp; prep)</th>
<th>smart</th>
<th>¬smart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>study</td>
<td>¬study</td>
</tr>
<tr>
<td>prepared</td>
<td>0.432</td>
<td>0.16</td>
</tr>
<tr>
<td>¬prepared</td>
<td>0.048</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Is *smart* independent of *study*?

Is *prepared* independent of *study*?
Exercise: Independence

<table>
<thead>
<tr>
<th></th>
<th>smart</th>
<th></th>
<th>¬smart</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>study</td>
<td>¬study</td>
<td>study</td>
<td>¬study</td>
</tr>
<tr>
<td>prepared</td>
<td>0.432</td>
<td>0.16</td>
<td>0.084</td>
<td>0.008</td>
</tr>
<tr>
<td>¬prepared</td>
<td>0.048</td>
<td>0.16</td>
<td>0.036</td>
<td>0.072</td>
</tr>
</tbody>
</table>

Is *smart* independent of *study*?

\[
P(\text{study} \land \text{smart}) = 0.432 + 0.048 = 0.48
\]

\[
P(\text{study}) = 0.432 + 0.048 + 0.084 + 0.036 = 0.6
\]

\[
P(\text{smart}) = 0.432 + 0.048 + 0.16 + 0.16 = 0.8
\]

\[
P(\text{study}) \times P(\text{smart}) = 0.6 \times 0.8 = 0.48
\]

So yes!

Is *prepared* independent of *study*?
Conditional Independence

• Absolute independence of $A$ and $B$:
  \[ A \perp B \iff P(A \land B) = P(A) \times P(B) \]
  \[ \iff P(A \mid B) = P(A) \]

\[ A \perp B \mid C \iff P(A \land B \mid C) = P(A \mid C) \times P(B \mid C) \]

• e.g., Moon-Phase and Burglary are conditionally independent given Light-Level

• This lets us decompose the joint distribution:
  \[ P(A \land B \land C) = P(A \mid C) \times P(B \mid C) \times P(C) \]
  
  – Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint
Take Home Exercise: Conditional independence

<table>
<thead>
<tr>
<th>P(smart ∧ study ∧ prep)</th>
<th>smart</th>
<th>¬smart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>study</td>
<td>¬study</td>
</tr>
<tr>
<td>prepared</td>
<td>0.432</td>
<td>0.16</td>
</tr>
<tr>
<td>¬prepared</td>
<td>0.048</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Is \textit{smart} conditionally independent of \textit{prepared}, given \textit{study}?

Is \textit{study} conditionally independent of \textit{prepared}, given \textit{smart}?
Summary: Essential Probability Concepts

• Marginalization:  
  \[ P(B) = \sum_{v \in \text{values}(A)} P(B \land A = v) \]

• Conditional Probability:  
  \[ P(A \mid B) = \frac{P(A \land B)}{P(B)} \]

• Bayes’ Rule:  
  \[ P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \]

• Independence:  
  \[ A \perp B \iff P(A \land B) = P(A) \times P(B) \]
  \[ \iff P(A \mid B) = P(A) \]
  \[ \iff P(A \land B \mid C) = P(A \mid C) \times P(B \mid C) \]
Density Estimation
How Can We Obtain a Joint Distribution?

**Option 1:** Elicit it from an expert human

**Option 2:** Build it up from simpler probabilistic facts

- e.g., if we knew

\[
P(a) = 0.7 \quad P(b | a) = 0.2 \quad P(b | \neg a) = 0.1
\]

then, we could compute \( P(a \land b) \)

**Option 3:** Learn it from data...

Based on slide by Andrew Moore
Learning a Joint Distribution

**Step 1:**
Build a JD table for your attributes in which the probabilities are unspecified.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

**Step 2:**
Then, fill in each row with:

\[
\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}
\]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Fraction of all records in which A and B are true but C is false.
Density Estimation

• Our joint distribution learner is an example of something called **Density Estimation**

• A Density Estimator learns a mapping from a set of attributes to a probability
Density Estimation

Compare it against the two other major kinds of models:

- Classifier
  - Prediction of categorical output
  - Input Attributes

- Density Estimator
  - Probability of inputs
  - Input Attributes

- Regressor
  - Prediction of real-valued output
  - Input Attributes
Evaluating Density Estimation

Test set criterion for estimating performance on future data

Input Attributes → Classifier → Prediction of categorical output

Input Attributes → Density Estimator → Probability

Input Attributes → Regressor → Prediction of real-valued output

Test set Accuracy

Test set Accuracy
Evaluating a Density Estimator

• Given a record \( x \), a density estimator \( M \) can tell you how likely the record is:

\[
\hat{P}(x \mid M)
\]

• The density estimator can also tell you how likely the dataset is:
  
  – Under the assumption that all records were independently generated from the Density Estimator’s JD (that is, i.i.d.)

\[
\hat{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_n \mid M) = \prod_{i=1}^{n} \hat{P}(x_i \mid M)
\]
Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

- 192 records in the training set

<table>
<thead>
<tr>
<th>mpg</th>
<th>modelyear</th>
<th>maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>75to78</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>europe</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>america</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>75to78</td>
<td>europe</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>europe</td>
</tr>
<tr>
<td>75to77</td>
<td>america</td>
<td>0.153061</td>
</tr>
<tr>
<td>asia</td>
<td>0.0255102</td>
<td></td>
</tr>
<tr>
<td>europe</td>
<td>0.0153061</td>
<td></td>
</tr>
<tr>
<td>78to83</td>
<td>america</td>
<td>0.0561224</td>
</tr>
<tr>
<td>asia</td>
<td>Never</td>
<td></td>
</tr>
<tr>
<td>europe</td>
<td>Never</td>
<td></td>
</tr>
<tr>
<td>75to77</td>
<td>america</td>
<td>0.0306122</td>
</tr>
<tr>
<td>asia</td>
<td>0.0408163</td>
<td></td>
</tr>
<tr>
<td>europe</td>
<td>0.0357143</td>
<td></td>
</tr>
<tr>
<td>78to83</td>
<td>america</td>
<td>0.112245</td>
</tr>
<tr>
<td>asia</td>
<td>0.0714286</td>
<td></td>
</tr>
<tr>
<td>europe</td>
<td>0.0357143</td>
<td></td>
</tr>
</tbody>
</table>

Slide by Andrew Moore
Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

• 192 records in the training set

<table>
<thead>
<tr>
<th>mpg</th>
<th>modelyear</th>
<th>maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>75to78</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>asia</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>79to83</td>
<td>america</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
<tr>
<td>good</td>
<td>75to78</td>
<td>europe</td>
</tr>
<tr>
<td>bad</td>
<td>75to78</td>
<td>europe</td>
</tr>
</tbody>
</table>

\[
\hat{P}(\text{dataset} \mid M) = \prod_{i=1}^{n} \hat{P}(x_i \mid M) = 3.4 \times 10^{-203} \quad \text{(in this case)}
\]
Log Probabilities

• For decent sized data sets, this product will underflow

\[ \hat{P}(\text{dataset} \mid M) = \prod_{i=1}^{n} \hat{P}(x_i \mid M) \]

• Therefore, since probabilities of datasets get so small, we usually use log probabilities

\[ \log \hat{P}(\text{dataset} \mid M) = \log \prod_{i=1}^{n} \hat{P}(x_i \mid M) = \sum_{i=1}^{n} \log \hat{P}(x_i \mid M) \]

Based on slide by Andrew Moore
Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

- 192 records in the training set

<table>
<thead>
<tr>
<th>mpg</th>
<th>modelyear</th>
<th>maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>75to78</td>
<td>asia</td>
</tr>
<tr>
<td>bad</td>
<td>70to74</td>
<td>america</td>
</tr>
</tbody>
</table>

\[
\log \hat{P}(\text{dataset} \mid M) = \sum_{i=1}^{n} \log \hat{P}(x_i \mid M)
\]

\[
= -466.19 \quad \text{(in this case)}
\]
Pros/Cons of the Joint Density Estimator

The Good News:

• We can learn a Density Estimator from data.

• Density estimators can do many good things...
  – Can sort the records by probability, and thus spot weird records (anomaly detection)
  – Can do inference
  – Ingredient for Bayes Classifiers (coming very soon...)

The Bad News:

• Density estimation by directly learning the joint is impractical, may result in adverse behavior
Curse of Dimensionality

$D = 1$

$D = 2$

$D = 3$
The Joint Density Estimator on a Test Set

<table>
<thead>
<tr>
<th>Set Size</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set 196</td>
<td>-466.1905</td>
</tr>
<tr>
<td>Test Set 196</td>
<td>-614.6157</td>
</tr>
</tbody>
</table>

- An independent test set with 196 cars has a much worse log-likelihood
  - Actually it’s a billion quintillion quintillion quintillion quintillion times less likely

- Density estimators can overfit...
  ...and the full joint density estimator is the overfittiest of them all!
Overfitting Density Estimators

If this ever happens, the joint PDE learns there are certain combinations that are impossible

\[
\log \hat{P}(\text{dataset} \mid M) = \sum_{i=1}^{n} \log \hat{P}(x_i \mid M)
\]

\[= -\infty \quad \text{if for any } i, \hat{P}(x_i \mid M) = 0\]