Bag of Words (cont)

Tamara Berg
CSE591
Recognizing People, Objects & Actions
Thursday, 2/9 – Talks in Wang Lecture Hall 1

3:00 Lecture 1: Stefanos Thomoplolous (concert pianist) – “The Use of Data Manipulation Techniques in the Music of Iannis Xenakis”

4:00 Lecture 2: Nicolas Maigret (from Art of Failure) - “Hello_World! : Global Data in the Works of Art of Failure”

5:15 Workshop Theatrical Performance “Phison 7: A Futuristic Look at Immigration through Interactive Performance

6:30 Art Exhibition Reception, SAC Gallery

Friday, 2/10 - Talks in Simons Center

4:00 Lecture 3: Rebecca Fiebrink (Princeton, Computer Science/Music) “Using End-User Machine Learning to Build Data-Driven Musical Instruments”

5:00 Lecture 4: Andreas Velten (MIT Media Lab) “Looking Around Corners: Capturing Light in Motion”

8:00 Concert - Staller
Paper Quiz - “Visual Categorization with Bags of Keypoints”

1) What is the high level goal of the paper?

2) Briefly describe a Naïve Bayes classifier and why it is a bag of features representation.

3) What other classification scheme did they try?

4) Which technique worked better?
Generative vs Discriminative

Discriminative version – build a classifier to discriminate between monkeys and non-monkeys.

\[ P(\text{monkey} \mid \text{image}) \]
Generative vs Discriminative

Generative version - build a model that generates images containing monkeys

$P(\text{image} | \text{monkey})$ vs $P(\text{image} | \text{not monkey})$
Generative vs Discriminative

Can use Bayes rule to compute $p(\text{monkey} | \text{image})$ if we know $p(\text{image} | \text{monkey})$

$$P(\text{monkey} | \text{image}) = P(\text{image} | \text{monkey}) \frac{P(\text{monkey})}{P(\text{image})}$$
Generative vs Discriminative

Can use Bayes rule to compute $p(\text{monkey} \mid \text{image})$ if we know $p(\text{image} \mid \text{monkey})$

$$P(\text{monkey} \mid \text{image}) = P(\text{image} \mid \text{monkey}) \frac{P(\text{monkey})}{P(\text{image})}$$

Discriminative

Generative
Talk Outline

1. Quick introduction to graphical models

2. Bag of words models
   - What are they?
   - Examples: Naïve Bayes, pLSA, LDA
Talk Outline

1. Quick introduction to graphical models

2. Bag of words models
   - What are they?
   - Examples: Naïve Bayes, pLSA, LDA
Probabilistic Models

- Models are descriptions of how (a portion of) the world works

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
Random Variables

Random variables

$$X = \{X_1, X_2, X_3, \ldots X_n\}$$

Let $x_i$ be a realization of $X_i$. 
Random Variables

Random variables $X = \{X_1, X_2, X_3, \ldots, X_n\}$

Let $x_i$ be a realization of $X_i$

A random variable is some aspect of the world about which we (may) have uncertainty.

Random variables can be:
- Binary (e.g. \{true, false\}, \{spam/ham\}),
- Take on a discrete set of values (e.g. \{Helicopter, Dolphin, Cougar, Laptop\}),
- Or be continuous (e.g. [0 1]).
Joint Probability Distribution

Random variables \[ X = \{X_1, X_2 X_3, \ldots X_n \} \]

Let \( x_i \) be a realization of \( X_i \)

Joint Probability Distribution:

\[ P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots X_n = x_n) \]

Also written \( p(x_1, x_2 x_3, \ldots x_n) \)

Gives a real value for all possible assignments.
Joint Probability Distribution:

\[ P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots X_n = x_n) \]

Also written \( p(x_1, x_2, x_3, \ldots, x_n) \)

Given a joint distribution, we can reason about unobserved variables given observations (evidence):

\[ P(X_q \mid x_{e1}, \ldots, x_{ek}) \]
Representation

Joint Probability Distribution:

\[ P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots X_n = x_n) \]

Also written \( p(x_1, x_2, x_3, \ldots x_n) \)

One way to represent the joint probability distribution for discrete \( X_i \) is as an n-dimensional table, each cell containing the probability for a setting of \( X \). This would have \( r^n \) entries if each \( X_i \) ranges over \( r \) values.
Joint Probability Distribution:

\[ P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots X_n = x_n) \]

Also written \( p(x_1, x_2 x_3, \ldots x_n) \)

One way to represent the joint probability distribution for discrete \( X_i \) is as an n-dimensional table, each cell containing the probability for a setting of \( X \). This would have \( r^n \) entries if each \( X_i \) ranges over \( r \) values.
Representation

Joint Probability Distribution:

\[ P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots X_n = x_n) \]

Also written \( p(x_1, x_2, x_3, \ldots x_n) \)

Graphical models represent joint probability distributions more economically, using a set of “local” relationships among variables.
Graphical Models

Graphical models offer several useful properties:

1. They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.

2. Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.

3. Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

from Chris Bishop
Main kinds of models

• Undirected (also called Markov Random Fields) - links express constraints between variables.

• Directed (also called Bayesian Networks) - have a notion of causality -- one can regard an arc from A to B as indicating that A "causes" B.
Main kinds of models

- Undirected (also called Markov Random Fields) - links express constraints between variables.

- Directed (also called Bayesian Networks) - have a notion of causality -- one can regard an arc from A to B as indicating that A "causes" B.
Directed Graphical Models

Directed Graph, \( G = (X,E) \)

Nodes \( X = \{X_1, X_2, X_3, \ldots X_n\} \)

Edges \( E = \{(X_i, X_j) : i \neq j\} \)

Each node is associated with a random variable
Directed Graphical Models

Directed Graph, \( G = (X,E) \)

Nodes \( X = \{X_1, X_2, X_3, \ldots X_n\} \)

Edges \( E = \{(X_i, X_j) : i \neq j\} \)

Each node is associated with a random variable

Definition of joint probability in a graphical model:

\[
p(x_1, \ldots x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})
\]

where \( \pi_i \) are the parents of \( x_i \)
Example
Example

Joint Probability:

\[ p(x_1, x_2 x_3, x_4, x_5, x_6) = \]

\[ p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2) p(x_5 \mid x_3) p(x_6 \mid x_2, x_5) \]
Example
Conditional Independence

Independence:

\[ p(x_A, x_B) = p(x_A) p(x_B) \]

Conditional Independence:

\[ p(x_A, x_C | x_B) = p(x_A | x_B) p(x_C | x_B) \]

Or,

\[ p(x_A | x_B, x_C) = p(x_A | x_B) \]
Conditional Independence

\[ p(x_1, x_2 x_3, x_4, x_5, x_6) = \]

\[ p(x_1) p(x_2 | x_1) p(x_3 | x_1, x_2) p(x_4 | x_1, x_2, x_3) p(x_5 | x_1, x_2, x_3, x_4) p(x_6 | x_1, x_2, x_3, x_4, x_5) \]
Conditional Independence

\[ p(x_1, x_2, x_3, x_4, x_5, x_6) = \]

\[ p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1, x_2) p(x_4 \mid x_1, x_2, x_3) p(x_5 \mid x_1, x_2, x_3, x_4) p(x_6 \mid x_1, x_2, x_3, x_4, x_5) \]

By Chain Rule (using the usual arithmetic ordering)
Example

Joint Probability:

\[ p(x_1, x_2, x_3, x_4, x_5, x_6) = \]
\[ p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) p(x_5 | x_3) p(x_6 | x_2, x_5) \]
Conditional Independence

\[ p(x_1, x_2, x_3, x_4, x_5, x_6) = \]

\[ p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1, x_2) p(x_4 \mid x_1, x_2, x_3) p(x_5 \mid x_1, x_2, x_3, x_4) p(x_6 \mid x_1, x_2, x_3, x_4, x_5) \]

By Chain Rule (using the usual arithmetic ordering)

Joint distribution from the example graph:

\[ p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2) p(x_5 \mid x_3) p(x_6 \mid x_2, x_5) \]
Conditional Independence

\[ p(x_1, x_2, x_3, x_4, x_5, x_6) = \]

\[ p(x_1) p(x_2 | x_1) p(x_3 | x_1, x_2) p(x_4 | x_1, x_2, x_3) p(x_5 | x_1, x_2, x_3, x_4) p(x_6 | x_1, x_2, x_3, x_4, x_5) \]

By Chain Rule (using the usual arithmetic ordering)

Joint distribution from the example graph:

\[ p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) p(x_5 | x_3) p(x_6 | x_2, x_5) \]

Missing variables in the local conditional probability functions correspond to missing edges in the underlying graph.

Removing an edge into node \( i \) eliminates an argument from the conditional probability factor \( p(x_i | x_1, x_2, ..., x_{i-1}) \)
Observations

• Graphs can have observed (shaded) and unobserved nodes. If nodes are always unobserved they are called hidden or latent variables.

• Probabilistic inference in graphical models is the problem of computing a conditional probability distribution over the values of some of the nodes (the “hidden” or “unobserved” nodes), given the values of other nodes (the “evidence” or “observed” nodes).
Inference – computing conditional probabilities

Conditional Probabilities: \[ p(x_1 | x_6) = \frac{p(x_1, x_6)}{p(x_6)} \]

Marginalization:
\[
p(x_{1},x_{6}) = \int \int \int \int p(x_{1}) p(x_{2} | x_{1}) p(x_{3} | x_{1}) p(x_{4} | x_{2}) p(x_{5} | x_{3}) p(x_{6} | x_{2}, x_{5})
\]
Inference Algorithms

• Exact algorithms
  – Elimination algorithm
  – Sum-product algorithm
  – Junction tree algorithm

• Sampling algorithms
  – Importance sampling
  – Markov chain Monte Carlo

• Variational algorithms
  – Mean field methods
  – Sum-product algorithm and variations
  – Semidefinite relaxations
1. Quick introduction to graphical models

2. Bag of words models
   - Reminder - What are they?
   - Examples: Naïve Bayes, pLSA, LDA
Bag of words for images

- Represent images as a “bag of words”
Exchangeability

• De Finetti Theorem of exchangeability (bag of words theorem): the joint probability distribution underlying the data is invariant to permutation.

\[ p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left( \prod_{i=1}^{N} p(x_i \mid \theta) \right) d\theta \]
Plates - "macro" that allows subgraphs to be replicated (graphical representation of the De Finetti theorem).
Talk Outline

1. Quick introduction to graphical models

2. Bag of words models
   - What are they?
   - Examples: Naïve Bayes, pLSA, LDA
A Simple Example – Naïve Bayes

We only specify (parameters):

$P(C)$ prior over class labels

$P(F_i | C)$ how each feature depends on the class

$P(C,F_1,F_2,...F_n) = P(C) \prod_{i} P(F_i | C)$
A Simple Example – Naïve Bayes

We only specify (parameters):

\[ P(C) \] prior over class labels
\[ P(F_i | C) \] how each feature depends on the class
A Spam Filter

- Naïve Bayes spam filter

- Data:
  - Collection of emails, labeled spam or ham
  - Note: someone has to hand label all this data!
  - Split into training, held-out, test sets

- Classifiers
  - Learn on the training set
  - (Tune it on a held-out set)
  - Test it on new emails

---

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ....

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Slide from Dan Klein
Example: Spam Filtering

- Model: \[ P(C, W_1 \ldots W_n) = P(C) \prod_i P(W_i|C) \]

- What are the parameters?
Example: Spam Filtering

- Model: \[ P(C, W_1 \ldots W_n) = P(C) \prod_i P(W_i|C) \]

- What are the parameters?

| \( P(C) \) | \( P(W|\text{spam}) \) | \( P(W|\text{ham}) \) |
|-----------|----------------|----------------|
| **ham**: 0.66 | the: 0.0156 to: 0.0153 and: 0.0115 of: 0.0095 you: 0.0093 a: 0.0086 with: 0.0080 from: 0.0075 ... | the: 0.0210 to: 0.0133 of: 0.0119 2002: 0.0110 with: 0.0108 from: 0.0107 and: 0.0105 a: 0.0100 ... |
| **spam**: 0.33 | ... | ... |

- Where do these tables come from?
Example: Spam Filtering

- **Model:** \[ P(C, W_1 \ldots W_n) = P(C) \prod_i P(W_i|C) \]

- **What are the parameters?**

| \( P(C) \) | \( P(W|\text{spam}) \) | \( P(W|\text{ham}) \) |
|-----------|----------------|----------------|
| ham: 0.66 | the: 0.0156    | the: 0.0210    |
| spam: 0.33| to: 0.0153     | to: 0.0133     |
|           | and: 0.0115    | and: 0.0105    |
|           | of: 0.0095     | of: 0.0119     |
|           | you: 0.0093    | 2002: 0.0110   |
|           | a: 0.0086      | with: 0.0108   |
|           | with: 0.0080   | from: 0.0107   |
|           | from: 0.0075   | a: 0.0100      |
|           | ...            | ...            |

- **Percentage of documents in training set labeled as spam/ham**

- **Where do these tables come from?**
Example: Spam Filtering

- Model: $P(C, W_1 \ldots W_n) = P(C) \prod_i P(W_i|C)$

- What are the parameters?

| $P(\text{C})$ | $P(W|\text{spam})$ | $P(W|\text{ham})$ |
|--------------|---------------------|---------------------|
| ham: 0.66    | the: 0.0156         | the: 0.0210         |
| spam: 0.33   | to: 0.0153          | to: 0.0133          |
|              | and: 0.0115         | of: 0.0119          |
|              | of: 0.0095          | 2002: 0.0110        |
|              | you: 0.0093         | with: 0.0108        |
|              | a: 0.0086           | from: 0.0107        |
|              | with: 0.0080        | and: 0.0105         |
|              | from: 0.0075        | a: 0.0100           |
|              | ...                 | ...                 |

In the documents labeled as spam, occurrence percentage of each word (e.g. # times “the” occurred/# total words).

- Where do these tables come from?
Example: Spam Filtering

- Model: $P(C, W_1 \ldots W_n) = P(C) \prod_i P(W_i|C)$

- What are the parameters?

| $P(C)$ | $P(W|spam)$ | $P(W|ham)$ |
|--------|-------------|-------------|
| ham : 0.66 | the : 0.0156 | the : 0.0210 |
| spam: 0.33 | to : 0.0153  | to : 0.0133 |
|          | and : 0.0115 | of : 0.0119 |
|          | of : 0.0095  | 2002: 0.0110|
|          | you : 0.0093 | with: 0.0108 |
|          | a : 0.0086   | from: 0.0107 |
|          | with: 0.0080 | and : 0.0105 |
|          | from: 0.0075 | a : 0.0100  |
|          | ...          | ...         |

In the documents labeled as ham, occurrence percentage of each word (e.g. # times “the” occurred/# total words).

- Where do these tables come from?
Classification

The class that maximizes:

\[
P(C, W_1, \ldots, W_n) = P(C) \prod_{i} P(W_i | C)
\]

\[
= \arg \max_{c \in C} P(c) \prod_{i} P(W_i | c)
\]
Classification

• In practice
  – Multiplying lots of small probabilities can result in floating point underflow
Classification

• In practice
  – Multiplying lots of small probabilities can result in floating point underflow
  – Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
Classification

• In practice
  – Multiplying lots of small probabilities can result in floating point underflow
  – Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
  – Since log is a monotonic function, the class with the highest score does not change.
Classification

• In practice
  – Multiplying lots of small probabilities can result in floating point underflow
  – Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
  – Since log is a monotonic function, the class with the highest score does not change.
  – So, what we usually compute in practice is:

$$c_{map} = \arg \max_{c \in C} \left[ \log P(c) + \sum_{i} \log P(W_i | c) \right]$$
Naïve Bayes on images
Visual Categorization with Bags of Keypoints
Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, Cédric Bray
Method

Steps:

– Detect and describe of image patches.
– Assign patch descriptors to a set of predetermined clusters (a visual *vocabulary*).
– Construct a *bag of keypoints, which counts the number of patches assigned* to each cluster.
– Apply a multi-class classifier (naïve Bayes), treating the bag of keypoints as the feature vector, and thus determine which category or categories to assign to the image.
Naïve Bayes

\[
P(C, F_1, F_2, \ldots, F_n) = P(C) \prod_{i} P(F_i | C)
\]

We only specify (parameters):

\[P(C)\] prior over class labels

\[P(F_i | C)\] how each feature depends on the class
Naive Bayes Parameters

Problem: Categorize images as one of 7 object classes using Naïve Bayes classifier:

– Classes: object categories (face, car, bicycle, etc)
– Features – Images represented as a histogram where bins are the cluster centers or visual word vocabulary. Features are vocabulary counts.

\[ P(C) \] treated as uniform.

\[ P(F_i | C) \] learned from training data – images labeled with category.
## Results

**Table 1.** Confusion matrix and the mean rank for the best vocabulary ($k=1000$).

<table>
<thead>
<tr>
<th>True classes</th>
<th>faces</th>
<th>buildings</th>
<th>trees</th>
<th>cars</th>
<th>phones</th>
<th>bikes</th>
<th>books</th>
</tr>
</thead>
<tbody>
<tr>
<td>faces</td>
<td>76</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>buildings</td>
<td>2</td>
<td>44</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>trees</td>
<td>3</td>
<td>2</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>cars</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>75</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>phones</td>
<td>9</td>
<td>15</td>
<td>1</td>
<td>16</td>
<td>70</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>bikes</td>
<td>2</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>8</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>books</td>
<td>4</td>
<td>19</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>69</td>
</tr>
<tr>
<td><strong>Mean ranks</strong></td>
<td><strong>1.49</strong></td>
<td><strong>1.88</strong></td>
<td><strong>1.33</strong></td>
<td><strong>1.33</strong></td>
<td><strong>1.63</strong></td>
<td><strong>1.57</strong></td>
<td><strong>1.57</strong></td>
</tr>
</tbody>
</table>