CS 688 Graphical Models, Spring 2017

Justin Domke

Agenda

1. **Typo Corrector**
2. It's all about the curse of dimensionality
3. Logistics
4. Prerequisites
5. What we will cover in the course

Typo corrector

Suppose we have a big database of $\leq T$ letter words:

```
duck
pile
mark
an**
dive
dog*

...

rug*
file
```

Typo corrector

Suppose we have a big database of $\leq T$ letter words:

```
duck
pile
mark
an**
dive
dog*

...

rug*
file
```

Our problem: we see new words, where 25% of the letters have been randomly corrupted.

```
frot
nice
v*he
qot*
vicn

...

taca
spix
```

Probabilistic model

Step one: Build a distribution $p(x)$ over all T-length sequences $x = (x_1, x_2, \ldots., x_T)$, each $x_t \in \{a, b, \ldots, *\}$

Probabilistic model

Step one: Build a distribution $p(x)$ over all T-length sequences $x = (x_1, x_2, \ldots, x_T)$, each $x_t \in \{a, b, \ldots, *\}$

$p(a, a, a, a) = .000001$

$p(a, a, a, b) = .000002$

...

$p(t, a, c, o) = .051231$

...

$p(*, *, *, *) = .00004$

Probabilistic model

Step two: Build a distribution $p(y|x)$ of "noisy" sequences y given "clean" sequences x.

Probabilistic model

Step two: Build a distribution $p(y|x)$ of "noisy" sequences y given "clean" sequences x.

What would this look like?

Probabilistic model

Step two: Build a distribution $p(y|x)$ of "noisy" sequences y given "clean" sequences x.

What would this look like?

$$p(y_t|x_t) = I(x_t = y_t) \times .75 + I(x_t \neq y_t) \times \frac{.25}{26}$$

Probabilistic model

Step two: Build a distribution $p(y|x)$ of "noisy" sequences y given "clean" sequences x.

What would this look like?

$$p(y_t|x_t) = I(x_t = y_t) \times .75 + I(x_t \neq y_t) \times \frac{.25}{26}$$

$$p(y|x) = \prod_{t=1}^{T} \left(I(x_t = y_t) \times .75 + I(x_t \neq y_t) \times \frac{.25}{26} \right)$$

Probabilistic model

Now, we have $p(x)$ (probability of a clean word) and $p(y|x)$ probability of a noisy word given a clean word.

Probabilistic model

Now, we have $p(x)$ (probability of a clean word) and $p(y|x)$ probability of a noisy word given a clean word.

Bayes' Equation tells us:

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}$$

Probabilistic model

Now, we have $p(x)$ (probability of a clean word) and $p(y|x)$ probability of a noisy word given a clean word.

Bayes' Equation tells us:

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}$$

For a given y could pick **most likely** x:

$$\arg\max_x p(x|y) = \arg\max_x p(x)p(y|x)$$

Probabilistic model

Now, we have $p(x)$ (probability of a clean word) and $p(y|x)$ probability of a noisy word given a clean word.

Bayes' Equation tells us:

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}$$

For a given y could pick **most likely** x:

$$\arg\max_x p(x|y) = \arg\max_x p(x)p(y|x)$$

But wait!

- How much time will this take?
- And how big does our dataset need to be?

Brute-force approach

```
For all x,
  Compute score(x) = p(x) p(y|x)
Return x with highest score
```

- How much time will the above algorithm take?
  - $O(27^T)$
- Is there a smarter algorithm?
  - No, not in general!
- How many free parameters does $p(x)$ have?
  - $27^T - 1$
- How many words do we need in our database to estimate these parameters reliably?
  - "A lot"

Brute-force approach

```
T    27^T
------------------------
1    27
2    729
3    19,683
4    531,441
5    14,348,907
6    387,420,489
7    10,460,353,203
8    282,429,536,481
9    7,625,597,484,987
10   205,891,132,094,649
```

Brute-force approach

```
T    27^T
------------------------
1    27
2    729
3    19,683
4    531,441
5    14,348,907
6    387,420,489
7    10,460,353,203
8    282,429,536,481
9    7,625,597,484,987
10   205,891,132,094,649
```

Lesson: For large T, we need more structure.

Agenda

1. Typo Corrector
2. **It's all about the curse of dimensionality**
3. Logistics
4. Prerequisites
5. What we will cover in the course

The curse of dimensionality

Graphical models assume that $p(x)$ can be written in a **factorized** form.

The curse of dimensionality

Graphical models assume that $p(x)$ can be written in a **factorized** form.

E.g.:

$$p(x_1, x_2, x_3, \ldots, x_9) = f(x_1, x_2) f(x_2, x_3) f(x_3, x_4) f(x_8, x_9)$$

The curse of dimensionality

Graphical models assume that $p(x)$ can be written in a **factorized** form.

E.g.:

$$p(x_1, x_2, x_3, \ldots, x_9) = f(x_1, x_2)f(x_2, x_3)f(x_3, x_4)f(x_8, x_9)$$

What does this buy us?

- Helps with the **statistical** curse of dimensionality.
- Helps with the **computational** curse of dimensionality.

The statistical curse of dimensionality

How many parameters does

$$p(x_1, x_2, \ldots, x_T)$$

have? (T variables, each with K values)

The statistical curse of dimensionality

How many parameters does

$$p(x_1, x_2, \ldots, x_T)$$

have? (T variables, each with K values)

- $K^T - 1$

The statistical curse of dimensionality

How many parameters does

$$p(x_1, x_2, \ldots, x_T)$$

have? (T variables, each with K values)

- $K^T - 1$

How many parameters does

$$p(x_1, x_2, x_3, \ldots, x_T) = f(x_1, x_2) f(x_2, x_3) \ldots$$

have?

The statistical curse of dimensionality

How many parameters does

$$p(x_1, x_2, \ldots, x_T)$$

have? (T variables, each with K values)

- $K^T - 1$

How many parameters does

$$p(x_1, x_2, x_3, \ldots, x_T) = f(x_1, x_2) f(x_2, x_3) \ldots$$

have?

- $(T - 1) \times (K^2 - 1)$

The statistical curse of dimensionality

First object of study in this course: The **representational capacity** of factorized models.

The statistical curse of dimensionality

First object of study in this course: The **representational capacity** of factorized models.

- **Question**: When can a probabilistic model be written in a factorized form?

The statistical curse of dimensionality

First object of study in this course: The **representational capacity** of factorized models.

- **Question**: When can a probabilistic model be written in a factorized form?
- **Answer**: When **conditional independencies** hold between the random variables

The statistical curse of dimensionality

First object of study in this course: The **representational capacity** of factorized models.

- **Question**: When can a probabilistic model be written in a factorized form?
- **Answer**: When **conditional independencies** hold between the random variables

We study two types of graphical models:

- **Directed** graphical models. (Bayesian networks, Markov models)
- **Undirected** graphical models. (Markov random fields, factor graphs)

The statistical curse of dimensionality

First object of study in this course: The **representational capacity** of factorized models.

- **Question**: When can a probabilistic model be written in a factorized form?
- **Answer**: When **conditional independencies** hold between the random variables

We study two types of graphical models:

- **Directed** graphical models. (Bayesian networks, Markov models)
- **Undirected** graphical models. (Markov random fields, factor graphs)

We understand **very precisely** how conditional indepdence assumptions can reduce the statistical curse of dimensionality.

We can cover this in a few lectures.

The computational curse of dimensionality.

Second object of study in the course: The **computational tractability** of factorized models.

The computational curse of dimensionality.

Second object of study in the course: The **computational tractability** of factorized models.

- If the graph is a tree (or close to it) we can often compute exact result using "message-passing" algorithms.

The computational curse of dimensionality.

Second object of study in the course: The **computational tractability** of factorized models.

- If the graph is a tree (or close to it) we can often compute exact result using "message-passing" algorithms.
- Otherwise, we typically need to rely on **approximate** algorithms.
  - Markov chain monte Carlo
  - Approximate message-passing algorithms
  - Variational methods

The computational curse of dimensionality

This is **subtle**.

- It is **not true** that just having a factorized model means we can do everything efficiently.
- We **don't** have a simple recipe for the best method to use in each case.
- **Bespoke** algorithms can make an enormous difference.

The computational curse of dimensionality

This is **subtle**.

- It is **not true** that just having a factorized model means we can do everything efficiently.
- We **don't** have a simple recipe for the best method to use in each case.
- **Bespoke** algorithms can make an enormous difference.

We will spend most of this class exploring these questions.

- What are the fundamental principles behind these different methods?
- When can we expect to be able to use graphical models efficiently?

Recapitulation: Why should you care about graphical models?

Probabilistic modeling is awesome.

- Natural and sometimes "optimal" way to solve problems.
- Can leverage lots of domain knowledge in setting up the problem.

Recapitulation: Why should you care about graphical models?

Probabilistic modeling is awesome.

- Natural and sometimes "optimal" way to solve problems.
- Can leverage lots of domain knowledge in setting up the problem.

Naive probabilistic modeling falls apart for high dimensions

- Too many parameters to estimate.
- Too much computational time needed.

Recapitulation: Why should you care about graphical models?

Probabilistic modeling is awesome.

- Natural and sometimes "optimal" way to solve problems.
- Can leverage lots of domain knowledge in setting up the problem.

Naive probabilistic modeling falls apart for high dimensions

- Too many parameters to estimate.
- Too much computational time needed.

Using a graphical model (factorized distribution) helps!

- Reduces number of parameters. (And we understand what we are assuming)
- Can help with computational issues. (But it's complicated!)

Agenda

1. Typo Corrector
2. It's all about the curse of dimensionality
3. **Logistics**
4. Prerequisites
5. What we will cover in the course

Logistics

Most important logistic: Please ask questions.

- I make mistakes!
- And otherwise… why are we in the same room?

Logistics

Most important logistic: Please ask questions.

- I make mistakes!
- And otherwise… why are we in the same room?

You do **not** need to have a super-specific technical question. **Vague** or **broad** questions are particularly encouraged.

Logistics

Most important logistic: Please ask questions.

- I make mistakes!
- And otherwise… why are we in the same room?

You do **not** need to have a super-specific technical question. **Vague** or **broad** questions are particularly encouraged.

The following are completely fine questions:

- "What's the point of learning this? How does it fit into the larger course?"
- "I feel like I'm missing the point of message-passing."
- "What does the c in x_c stand for again?"

Logistics

- What: This course: CS688 Graphical Models
- Where: CS 142
- When: Tuesday and Thursday 1:00pm to 2:15 pm.
- Who:
  - Instructor: Justin Domke
  - TA: Hang Su
- Office Hours: TBD
- URL: `http://people.cs.umass.edu/~domke/courses/compsci688/`
- Textbook: Kevin Murphy's *Machine Learning: a Probabilistic Perspective*

Grades

- Homework Assignments: 50%
- Final Exam: 30%
- Quizzes: 15%
- Participation: 5% (including online)

How to contact us

All questions should be done through **Piazza**:

- You should already be enrolled.

This allows for great **knowledge sharing**.

- Post questions for the class, not just for us!
- Answer any questions.

How to contact us

All questions should be done through **Piazza**:

- You should already be enrolled.

This allows for great **knowledge sharing**.

- Post questions for the class, not just for us!
- Answer any questions.

What happens if you email us?

1. We reply "please post to piazza". ☹️
2. We feel guilty. ☹️
3. You feel annoyed. ☹️

When and where would you like office hours?

- Monday morning / afternoon
- Tuesday before / after class
- Wednesday morning / afternoon
- Thursday before / after class
- Friday morning / afternoon

There is a poll on Piazza!

Agenda

1. Typo Corrector
2. It's all about the curse of dimensionality
3. Logistics
4. **Prerequisites**
5. What we will cover in the course

Prerequisites

Question: Will this class focus on math?

Prerequisites

Question: Will this class focus on math?

Short answer Yes.

Prerequisites

Question: Will this class focus on math?

Short answer Yes.

Long answer Yyyyyyyeeeeeeeeessssssss!!!!

Prerequisites

Question: Will this class focus on math?

Short answer Yes.

Long answer Yyyyyyyeeeeeeeeeesssssssss!!!!

Why?

- Machine learning *is* applied math.
- The goal of this course is to give you the foundations to do AI research.
- Goal is not how to use existing tools, but understand how to combine them and invent new ones.

Prerequisites

From course webpage:

> I will assume a strong working knowledge of Linear Algebra, Probability theory, programming ability in some language (e.g. Python) and that you have some familiarity with basic machine learning methods.

Prerequisites

From course webpage:

> I will assume a strong working knowledge of Linear Algebra, Probability theory, programming ability in some language (e.g. Python) and that you have some familiarity with basic machine learning methods.

Take this seriously!

- Past experience with talented and hard-working students but weak background: Struggle in this course.

Example Questions (Probability)

Given a fair six-sided die, what is more probable?

A) Rolling

`1,6,2,6,1`

or

B) Rolling

`2,2,2,2,2`

?

Example Questions (Probability)

Given a fair six-sided die, what is more probable?

A) Rolling one 2 and four 3s (in any order)

or

B) Rolling five twos?

Example Questions (Probability)

Suppose that A and B are binary random variables.

$$P(A = 1) = 1/2$$

$$P(A = 0) = 1/2$$

$$P(B = A) = 3/4$$

What is

$$P(B = 1)?$$

Example Questions (Probability)

Suppose that A and B are binary random variables.

$$P(A = 1) = 1/2$$

$$P(A = 0) = 1/2$$

$$P(B = A) = 3/4$$

What is

$$P(A = 1|B = 1)?$$

Example Questions (Linear Algebra)

Suppose that

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Can you name an eigenvector of A?

Example Questions (Linear Algebra)

Suppose that

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

What is

$$x^T A x?$$

Agenda

1. Typo Corrector
2. It's all about the curse of dimensionality
3. Logistics
4. Prerequisites
5. **What we will cover in the course**

What we will cover

- Directed models
    - Bayesian Networks
- Undirected models
    - Markov Random Fields
    - Conditional Random Fields
- Maximum likelihood learning
    - Optimization
- Exact Inference: Message-Passing
- Approximate Inference: Variational Inference
- Approximate Inference: Markov Chain Monte Carlo Methods

Requests?

Applications

- Speech recognition
- Image recognition and labeling
- Image modeling
- Action recognition
- Modeling sensor networks
- Social network analysis
- Recommender systems
- Evolutionary biology
- Proteomics and Genomics
- Medical decision making
- Information extraction
- Text modeling
- Bayesian statistics

Upcoming

- Reading:
  - Your favorite linear algebra text
  - Chapter 1 (or your favorite machine learning text)
  - Sections 2.1-2.5
- Thursday: "Math Camp" taught by your TA, Hang Su
- Next Tuesday: For credit quiz on prerequisites, in class.

Shameless request

Are you familiar with these things?

- HTML/Javascript latex renderers
  - Mathjax
  - Katex
- Markdown/Javascript based slides
  - Remark.js
  - React.js