Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 3: A Linear view of Discriminative and Generative Models

January 28, 2025

Reading: See course homepage

Logistics Review

- Lecture scribe sign-up sheet
- Instructor: Ben Lengerich
  - Office Hours: **Thursday 3:30-4:30pm**, 7278 Medical Sciences Center
  - Email: lengerich@wisc.edu
- TA: Chenyang Jiang
  - Office Hours: Monday 11am-12pm, 1219 Medical Sciences Center
  - Email: cjiang77@wisc.edu

Homework 1

- Due Friday at midnight.
- Submit via **Canvas**.
- How is it going?

Questions about Course Logistics?

Today

- Introduction to LaTeX
- Discriminative vs Generative Models:
  - Logistic Regression (Discriminative)
  - Naïve Bayes (Generative)

Introduction to LaTeX

What is LaTeX?

- LaTeX is a document preparation system and markup language used to create high-quality documents
- Key ideas:
  - Separation of content and formatting.
  - Not WYSIWYG (like Word); instead uses a language for precise design.
- Implications:
  - Handles complex layouts like tables, equations, and citations seamlessly.
  - Some upfront work, but no fighting the system for layout.

```latex
\begin{solution}
        Write your solution here. For multiple choice
        questions, only the letter answer is required.
    \begin{parts}
        \part Solution for (a)
        \part Solution for (b)
        \part Solution for (c)
        \part Solution for (d)
    \end{parts}
\end{solution}
```

Answer: Write your solution here.

(a) Solution for (a)

(b) Solution for (b)

(c) Solution for (c)

(d) Solution for (d)

Why use LaTeX?

- **Professional Output**: Superior typesetting for academic papers, theses, and technical reports.
- **Handles Complexities**: Math **formulas**, **bibliographies**, and **cross-referencing** are simple and accurate.
- **Automation**: Automatically generates **tables of contents**, **indices**, and **citations**.
- **Reproducibility**: Documents look consistent regardless of the platform.

Basics of a LaTex document

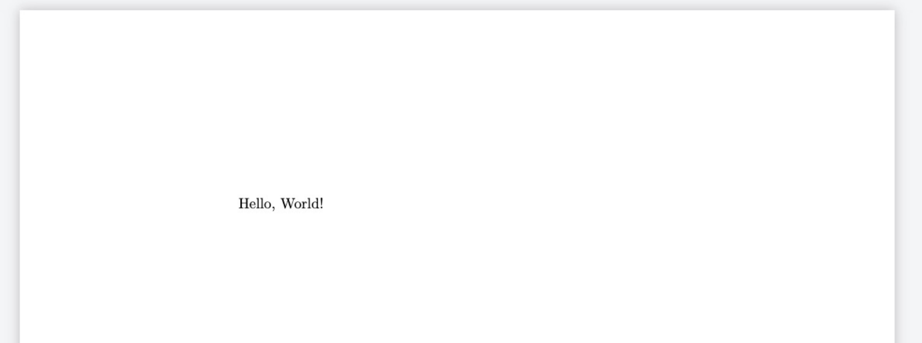
- Minimal example of a LaTeX document:

Preamble

Content

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\begin{document}
Hello, World!
\end{document}
```



Essential Commands

- **Text Formatting**:
  - Bold: `\textbf{bold text}`
  - Italics: `\textit{italic text}`
  - Underline: `\underline{text}`
- **Lists:**

```
\begin{itemize}
    \item Item 1
    \item Item 2
\end{itemize}
```

Math in LaTeX

- **In-line math:**
  - Single $ to start/end math environment: `$E = mc^2$`
- **Block math:**

```
\[
E = mc^2
\]
```

Figures

Preamble

```
\usepackage{graphicx}


\begin{figure}[h]
   \includegraphics[width=0.5\linewidth]{image.png}
   \caption{A sample figure}
\end{figure}
```

Where to put the figure

A caption for the figure

Define the width of the figure

Your image

Tables

How many columns and to arrange the column text (c = center)

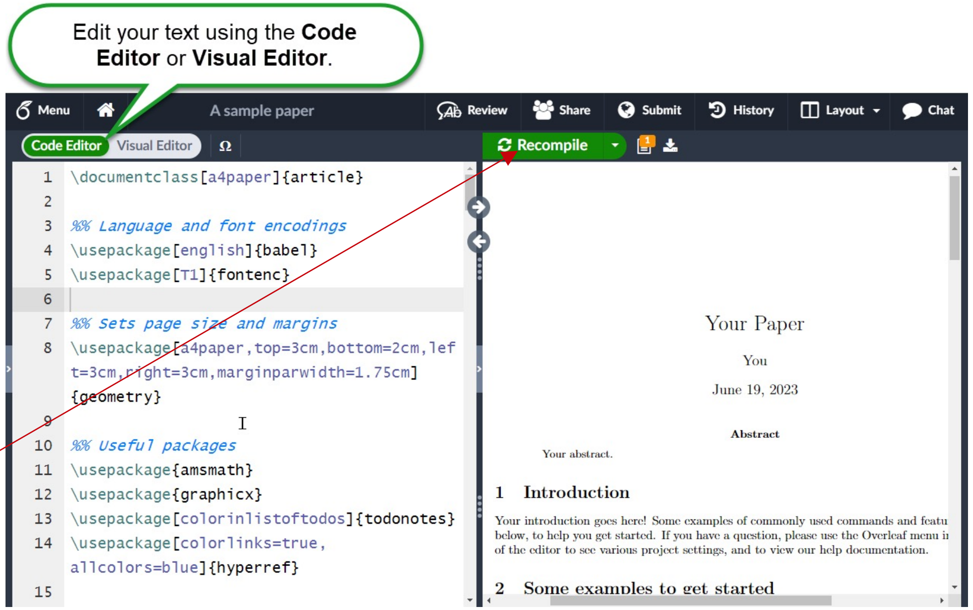
Start table

Horizontal line

```
\begin{tabular}{|c|c|}
   \hline
   Column 1 & Column 2 \\
   \hline
   Data 1 & Data 2 \\
   \hline
\end{tabular}
```

Where to use LaTeX

- I recommend [Overleaf]
  - Live collaboration tools
  - Integrates with Dropbox, etc for automatic backups
- Also possible to use local editors: TeXShop, MikTeX, TeXworks
- Workflow:
  - Write LaTeX code.
  - Compile:
    - Use the "Recompile" button on Overleaf
    - Use `pdflatex` or `xelatex` locally.
  - View and iterate on the generated PDF.



Resources

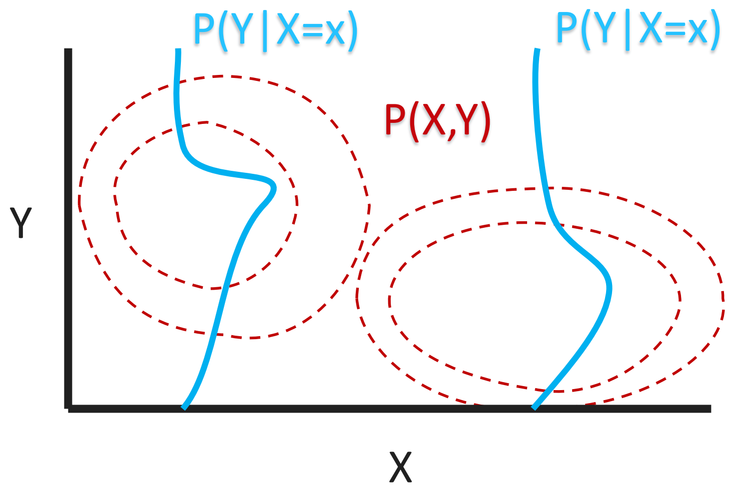
- Overleaf Tutorials
- LaTeX Wikibook
- TeX Stack Exchange for Q&A.

Questions about LaTeX?

A Linear View of Discriminative and Generative Models

Generative and Discriminative Models

- **Generative**:
  - Models the joint distribution $\mathrm{P}(X, Y)$.
- **Discriminative**:
  - Models the conditional distribution $\mathrm{P}(Y|X)$.



Two paths to P(Y|X)

- **Discriminative**:

Observe X, Y → Learn P(Y|X)

- **Generative**:
  - Learn $\mathrm{P}(X|Y), \mathrm{P(Y)}$
  - Calculate $P(X) = \int_Y P(X,Y)dY$

Observe X, Y

$$\mathrm{P(Y|X)} = \frac{\mathrm{P(X|Y)P(Y)}}{P(X)}$$

Two paths to classification

- **Discriminative**:

Observe X, Y $\rightarrow$ $\hat{Y} = \mathrm{argmax}_Y \mathrm{P}(\mathrm{Y}|\mathrm{X})$

- **Generative**:
  - Learn $\mathrm{P}(X|Y), \mathrm{P}(\mathrm{Y})$
  - ~~Calculate $P(X) = \int_Y P(X,Y)dY$~~

Observe X, Y

$\hat{Y} = \mathrm{argmax}_Y \mathrm{P}(\mathrm{X}|\mathrm{Y})\mathrm{P}(\mathrm{Y})$

Example Discriminative Model: Logistic Regression

Observe X, Y → Learn P(Y|X)

- Parameterize:
  - $P(Y=1|X) = \sigma(\theta^T X)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.
  - $P(Y=0|X) = 1 - P(Y=1|X)$
  - **Why this parameterization?**

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = \log \frac{\sigma(\theta^T X)}{1-\sigma(\theta^T X)}$$

$$= \log \frac{\frac{1}{1+e^{-\theta^T X}}}{1-\frac{1}{1+e^{-\theta^T X}}} = \log \frac{\frac{1}{1+e^{-\theta^T X}}}{\frac{(1+e^{-\theta^T X})-1}{1+e^{-\theta^T X}}} = \log \frac{\frac{1}{1+e^{-\theta^T X}}}{\frac{e^{-\theta^T X}}{1+e^{-\theta^T X}}}$$

$$= \log \frac{1}{e^{-\theta^T X}} = \log e^{\theta^T X} = \boldsymbol{\theta^T X}$$

Example Discriminative Model: Logistic Regression

Observe X, Y → Learn P(Y|X)

- Parameterize:
  - $P(Y=1|X) = \sigma(\theta^T X)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.
  - $P(Y=0|X) = 1 - P(Y=1|X)$
- Estimate $\hat{\theta}$ from observations:
  - $\hat{\theta} = \text{argmax}_\theta \prod_i P(Y_i|X_i;\theta)$
    $= \text{argmax}_\theta \sum_i [Y_i \log \sigma(\theta^T X_i) + (1-Y_i)\log(1 - \sigma(\theta^T X_i))]$
- Calculate $P(Y=1|X) = \sigma(\theta^T X)$

Example Generative Model: Naïve Bayes

Learn $\mathrm{P}(X|Y), \mathrm{P}(\mathrm{Y})$

Observe X, Y

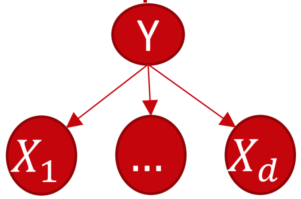
$$\mathrm{P(Y|X)} = \frac{\mathrm{P(X|Y)P(Y)}}{P(X)}$$

- Parameterize:
  - Assume $P(X|Y) = \prod_{j=1}^{d} P(X_j|Y)$,
  - $P(X_j|Y) = N(\mu_{jk}, \sigma_{jk}^2)$

Conditional independences of features X | Y

$$P(Y = k) = \frac{\#\ of\ samples\ with\ Y{=}k}{Total\ samples}$$

Frequency of labels



Example Generative Model: Naïve Bayes

Learn $\mathrm{P}(X|Y), \mathrm{P}(\mathrm{Y})$

Observe X, Y

$$\mathrm{P}(\mathrm{Y}|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\mathrm{Y})\mathrm{P}(\mathrm{Y})}{P(X)}$$

- Parameterize:
  - Assume $P(X|Y) = \prod_{j=1}^{d} P\left(X_j \middle| Y\right)$,

$$P(Y = k) = \frac{\#\ of\ samples\ with\ Y{=}k}{Total\ samples}$$

- Estimate:
  - $\hat{\mu}, \hat{\sigma} = \mathrm{argmax}_{\mu,\sigma} P(X|Y)$
- Calculate $P(Y = 1|X) = \frac{\prod_{j=1}^{d} P\left(X_j \middle| Y = 1\right) P(Y=1)}{P(X)}$

Summary

- **Discriminative**:

Observe X, Y $\longrightarrow$ Learn P(Y|X)

- **Generative**:
  - Learn $\mathrm{P}(X|Y), \mathrm{P}(\mathrm{Y})$
  - Calculate $P(X) = \int_Y P(X, Y) dY$

Observe X, Y

$$\mathrm{P}(\mathrm{Y}|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\mathrm{Y})\mathrm{P}(\mathrm{Y})}{P(X)}$$

What about MAP / Regularization?

Prior $\mathrm{P}(\theta)$

Logistic Regression:

Observe X, Y → Learn $\mathrm{P}(\mathrm{Y}|\mathrm{X};\theta)$

- Parameterize:
  - $P(Y=1|X)=\sigma(\theta^T X)$, where $\sigma(z)=\frac{1}{1+e^{-z}}$ is the sigmoid function.
  - $P(Y=0|X)=1-P(Y=1|X)$
- Estimate $\hat{\theta}$ from observations:
  - $\hat{\theta}=\mathrm{argmax}_\theta \prod_i P(Y_i|X_i;\theta)P(\theta)$
    $=\mathrm{argmax}_\theta \sum_i[Y_i \log\sigma(\theta^T X_i)+(1-Y_i)\log(1-\sigma(\theta^T X_i))]-R(\theta)$
- Calculate $P(Y|X)$

Discriminative vs Generative Models

- Discriminative models optimize the conditional likelihood:

$$\widehat{\theta_{disc}} = \operatorname{argmax}_{\theta} P(Y|X;\theta)$$

- Generative models optimize the joint likelihood:

$$\widehat{\theta_{gen}} = \operatorname{argmax}_{\theta} P(X,Y;\theta)$$

Are these the same optimization?

Discriminative vs Generative Models

- Discriminative models optimize the conditional likelihood:

$$\widehat{\theta_{disc}} = \text{argmax}_\theta P(Y|X;\theta) = \text{argmax}_\theta \frac{P(X|Y;\theta)P(Y;\theta)}{P(X;\theta)}$$

- Generative models optimize the joint likelihood:

$$\widehat{\theta_{gen}} = \text{argmax}_\theta P(X,Y;\theta) = \text{argmax}_\theta P(X|Y;\theta)P(Y;\theta)$$

Are these the same optimization?

Same optimization when $P(X;\theta)$ is invariant to θ

Logistic Regression vs Naïve Bayes

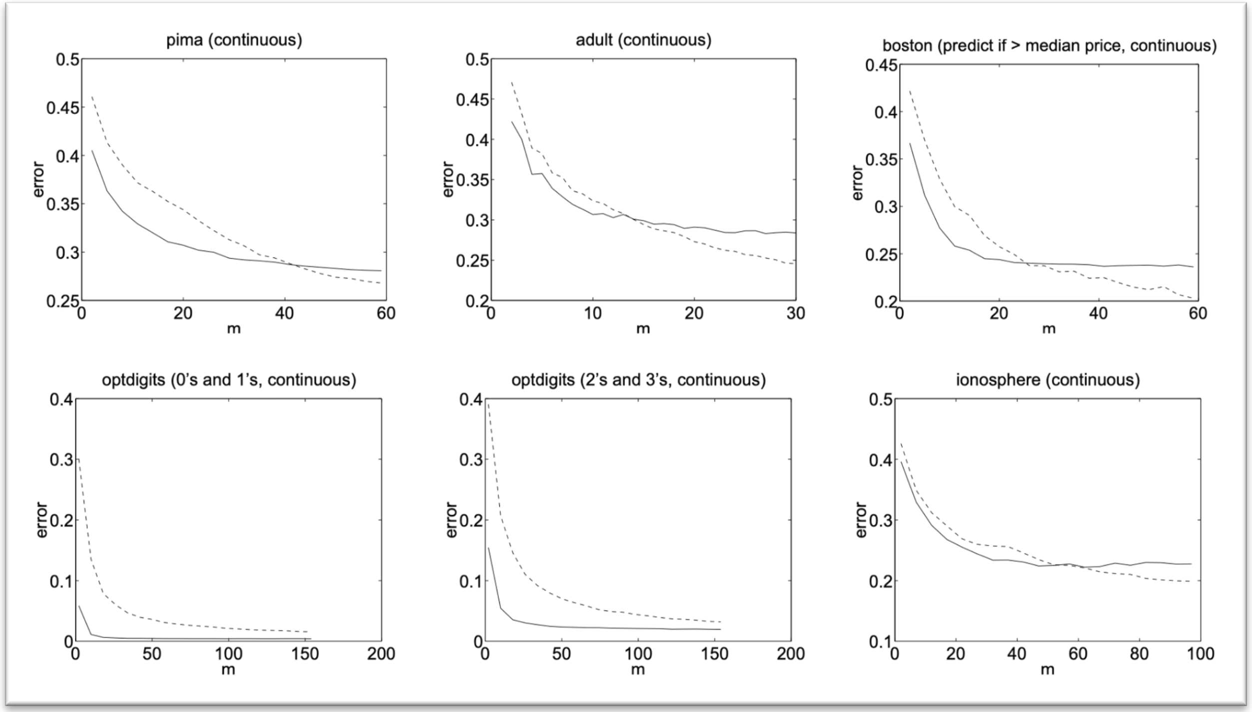
Logistic Regression	Naïve Bayes
Discriminative	Generative
Defines $P(Y\|X;\theta)$	Defines $P(X,Y;\theta)$
Estimates $\widehat{\theta_{lr}} = \text{argmax}_\theta P(Y\|X;\theta)$	Estimates $\widehat{\theta_{nb}} = \text{argmax}_\theta P(X,Y,\theta)$
Lower asymptotic error on classification	Higher asymptotic error on classification
Slower convergence in terms of samples	Faster convergence in terms of samples

Andrew Ng's Insight

- "While discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster."

Why?

········ LR
—— NB



Ng & Jordan 2001

Andrew Ng's Insight

- "While discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster."
- Underlying assumption of this statement:
  - Generative models of the form $\boldsymbol{P(X, Y, \theta)}$ make **more simplifying assumptions** than do discriminative models of the form $\boldsymbol{P(Y|X, \theta)}$**.**
  - **Not always true**
  - "So far there is no theoretically correct, general criterion for choosing between the discriminative and the generative approaches to classification of an observation $\mathbf{x}$ into a class y; the choice depends on the relative confidence we have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(\mathbf{x}, y)$ for the data."

Xue & Tittering 2008

Modern Deep Generative Models

January 27, 2025:

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON