UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Spring 2025)

Machine Learning

Lecture 12: Probabilistic Models

Chong Liu

Department of Computer Science

Mar 26, 2025

Today

- Lectures after midterm
- Discriminative model vs generative model
- Maximum likelihood estimation
  - Linear regression
- Naïve Bayes model

Lectures after midterm

- Lecture 12: Probabilistic models in supervised learning
  - Maximum likelihood estimation
  - Naïve Bayes model
- Lecture 13-16: Advanced techniques to improve supervised learning
  - Theoretical foundation: Error decomposition
  - Ensemble method (combining multiple classifiers)
  - Kernel method (feature transformation)
  - Non-linear model (neural network)
- Lecture 17-18: Unsupervised learning
  - Clustering
  - Dimension reduction
- Lecture 19: Advanced topic: Decision making
- Lecture 20: Course review

So far we have learned a lot about ML, but...

- We learned how to
  - Specify a hypothesis class
  - Work out the possible shapes of decision boundaries
  - Train a model by solving an optimization problem

- How did we come up with the hypothesis classes in the first place?
  - We brainstormed… and used
    1. decision trees
    2. linear-classifiers, thresholding a weighted linear combination of features.
  - But how do we know the resulting decision boundaries are appropriate for the problems we hope to solve?

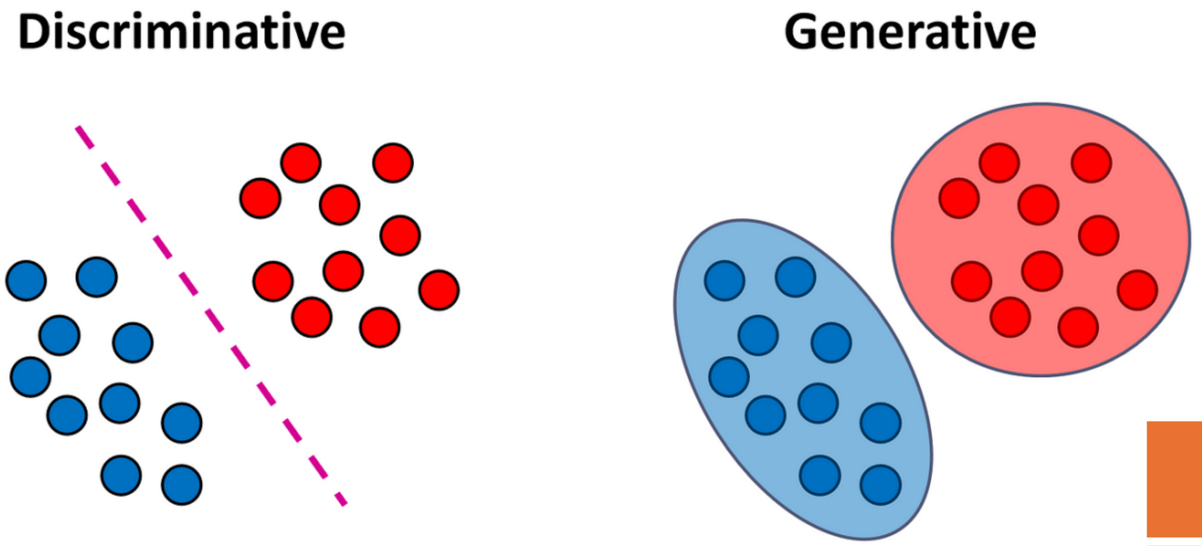
We learned about directly modelling the predictive functions. There is another way… called "Probabilistic modelling"

- We can model how the data is generated in the first place.

  - Model the labeling process via a **conditional distribution** $P(y|x)$.
  - ***(Probabilistic) discriminative model***
    - Non-probabilistic discriminative models:
      - Decision-trees
      - Linear classifiers

  - Model the **joint distribution** $P(x, y)$ by modeling the label distribution $P(y)$ and a generative process $P(x|y)$.
  - ***Generative model.***

Discriminative models vs generative models



	Non-probabilistic	Probabilistic
Discriminative model (how data can be separated?)	Modeling predictive function	Modeling $P(y\|x)$
Generative model (How data is generated?)		Modeling $P(x,y)$ by label distribution $P(y)$ and generative process $P(x\|y)$

Image Credit: Dr. Roi Yehoshua

Probabilistic modelling

- Hard prediction:
  - $h(x) = \mathrm{argmax}_y\ P(y|x)$
  - "Bayes optimal":
    - If the label generative process is indeed $P(y|x)$

- Soft prediction $P(y|x)$
  - Quantifying uncertainty
  - More informative than the score function
  - More interpretable / explainable

How to model $P(y|x)$?

- Bayes rule:
  - $P(y|x) = \frac{P(y)P(x|y)}{P(x)}$
- Key idea:
  - $\text{argmax}_y\, P(y|x) \leftrightarrow \text{argmax}_y\, P(y)P(x|y)$
  - Why?
    - y doesn't depend on $P(x)$.
  - $P(y)$ (distribution of label):
    - Can be estimated by counting labels in training set.
  - $P(x|y)$ (data generating process)

Directly modeling $P(x|y)$ is challenging

Binary vectors, 2^3 rows + binary output $Y \in \{0, 1\}$

x_1	x_2	x_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

1. What is the number of data points needed to estimate $p(x|y)$?

2. What happens if there are d binary features?

We need maximum likelihood estimation!

Maximum likelihood estimation

- Used since Gauss, Laplace, etc….
- Popularized / carefully analyzed by Ronald Fisher.

- Which distribution is more *likely* to have produced the data?

$$\max_{P \in \Pi} f_{\mathrm{Data} \sim P}(\mathrm{Data})$$

What is the difference between **probability** and **likelihood**?

- P(Data; Parameter)
  - If it is a function of the data, then it's probability.
  - If it is a function of the parameter while the data is fixed, then it is likelihood.

Recap: Estimating the mean of Gaussian distribution

- Data

$$X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- Likelihood:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- The MLE problem:

$$\hat{\mu} = \arg\max_{\mu\in[0,1]} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

MLE for linear regression

- $P(y|x)$ is modeled by "Linear Gaussian model"

$$y_i = x_i^T \theta^* + \epsilon_i \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

- Data:

$$(x_1, y_1), ..., (x_n, y_n)$$

- In-class exercise: Work out the optimization problem to solve for the MLE for θ^*.

After we fit the MLE, how to make predictions? The idea is to just "Plug-In"

- For classification problems

$$h^*(x) = \max_y p_\theta(y|x) \xrightarrow{\text{Plug in}} \hat{h}(x) = \max_y p_{\hat{\theta}}(y|x)$$

- For regression problems

$$h^*(x) = \mathbb{E}_\theta[y|x] \xrightarrow{\text{Plug in}} \hat{h}(x) = \mathbb{E}_{\hat{\theta}}[y|x]$$

  - Linear Gaussian model: $y \sim N(x^{\mathrm{T}}\hat{\theta}, \sigma^2)$