UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Fall 2024)

Machine Learning

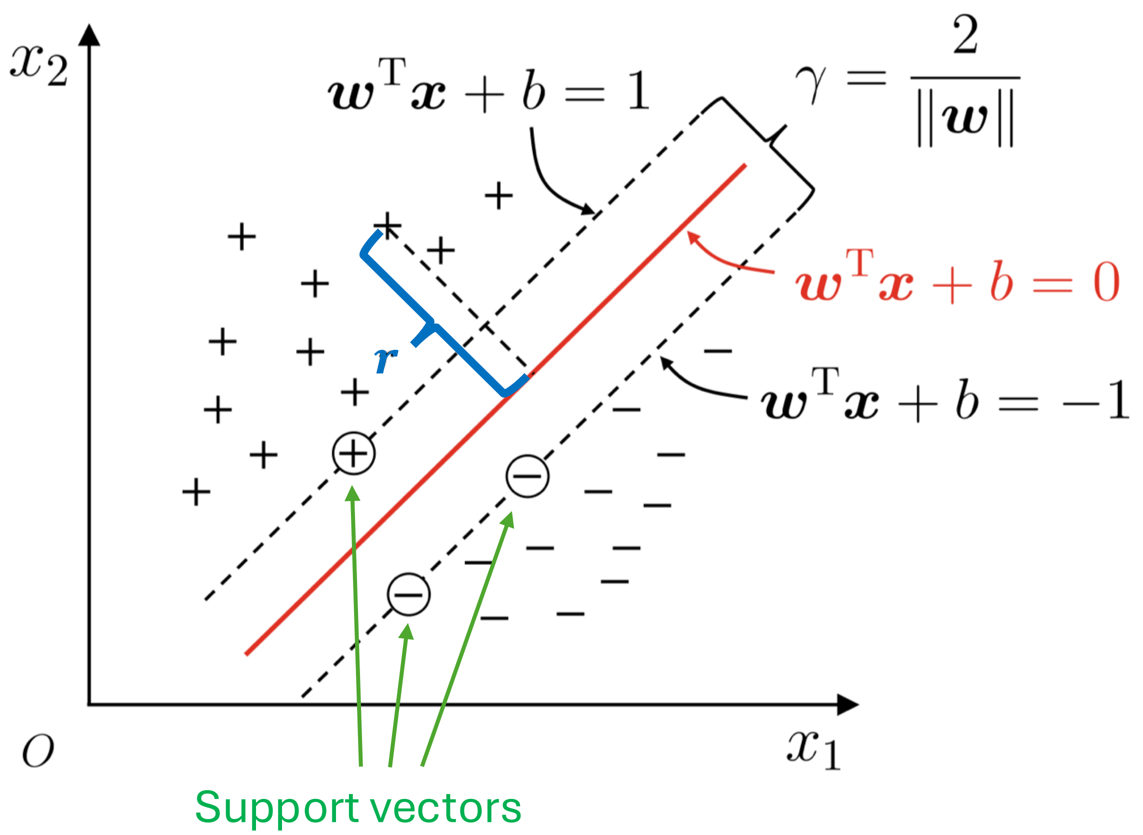
Lecture 12: Maximum Likelihood Estimation

Chong Liu
Assistant Professor of Computer Science

Oct 8, 2024

Recap: Support Vector Machines (SVM)

- Key idea of SVM:
  - If $y = 1, w^T x + b \geq 1$
  - If $y = -1, w^T x + b \leq -1$
- Total margin between support vectors:
  - $\gamma = \frac{2}{||w||}$
- Optimization problem of SVM:
  - $\max_{w,b} \frac{2}{||w||}$
  - s.t. $y_i(w^T x_i + b) \geq 1, i = 1, \dots, n$



Recap: Philosophy of designing ML algorithms

- Regularization:
  - Control the complexity of parameters
  - Prevent overfitting
  - Fun fact: L-2 regularization is associated with max margin classifier
- Optimization
  - Toolbox of ML
  - ML problem => optimization problem
    - Direct solver, GD, SGD, and much more!
  - Minimize the loss / parameter complexity / soft-margin tolerance
  - Maximize the margin

Today

- Discriminative model vs generative model

- Maximum likelihood estimation
    - Linear regression
    - Logistic regression

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?

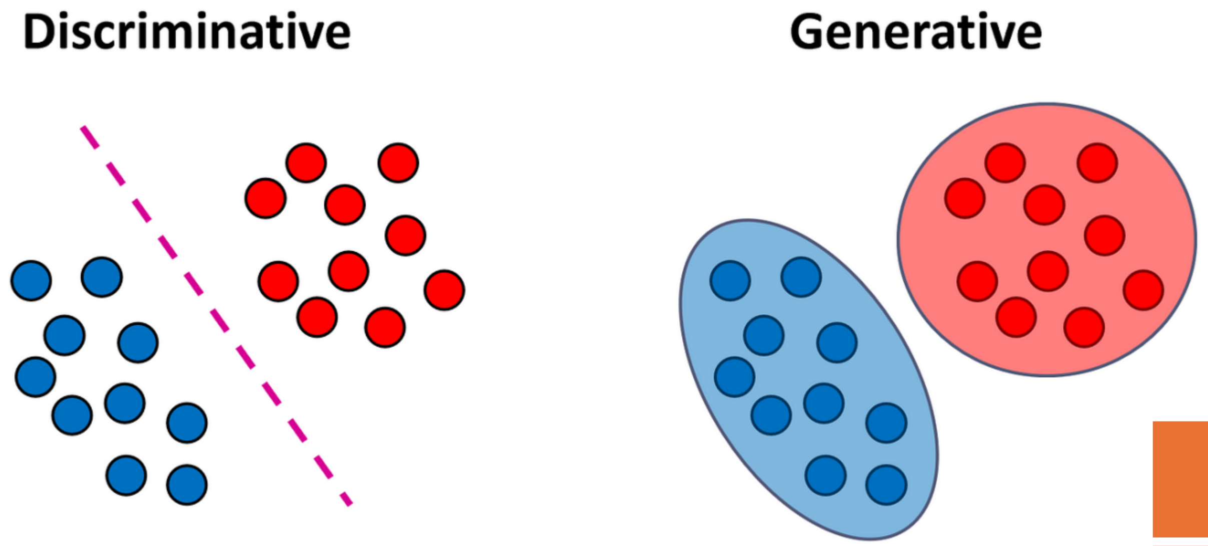
We learned about directly modelling the predictive functions. There is another way… called “Probabilistic modelling”

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

  - Model the labeling process via a **conditional distribution** $P(y|x)$**.** This is known as a ***(probabilistic) discriminative model***.
    - Specifying decision-trees / linear classifiers / shapes of decision boundaries should be considered non-probabilistic discriminative models.

  - Model the **joint distribution** $P(x, y)$**.** Often one models the label distribution $P(y)$ and a generative process $P(x|y)$**.** This is known as a ***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 $Y \in \{0, 1\}$

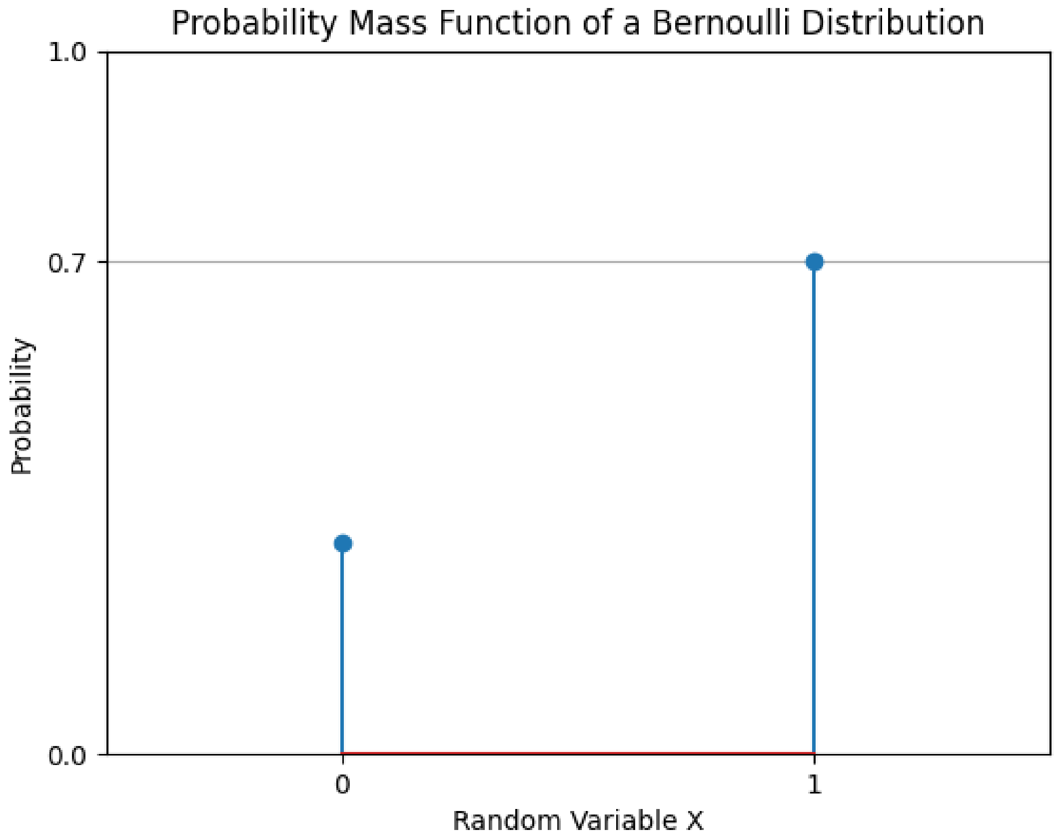
x_1	x_2	x_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

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

2. What happens if there are d binary features?

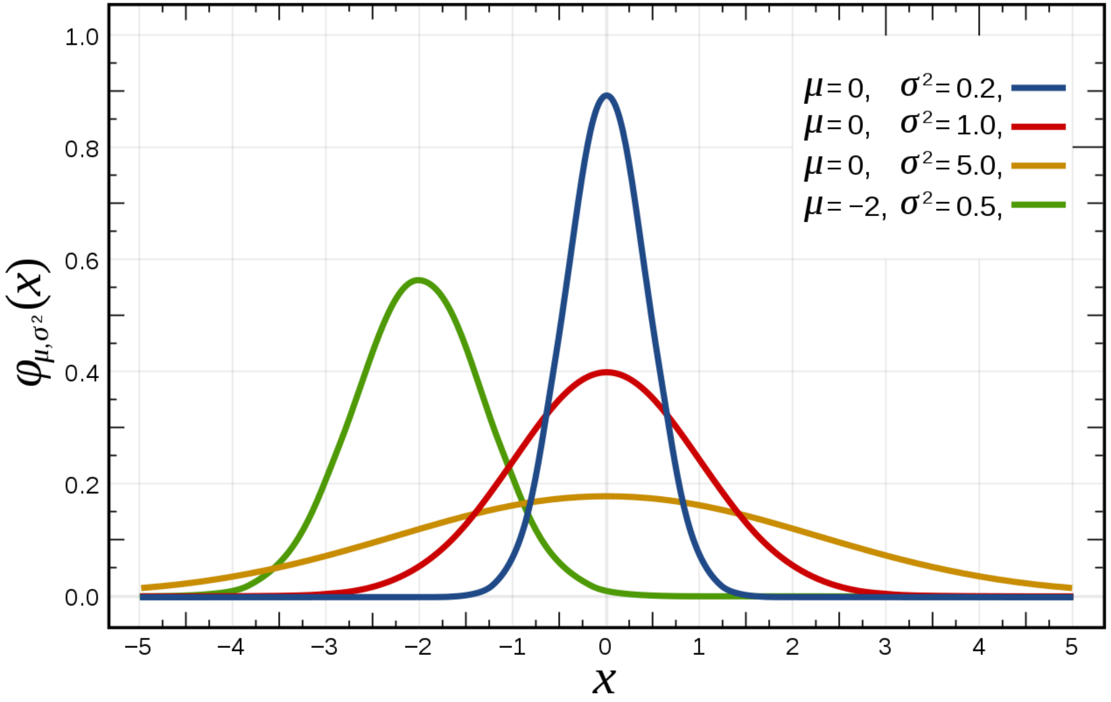
We need maximum likelihood estimation!

Recap: Bernoulli Distribution $X \sim \mathrm{Ber}(p)$



$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

Recap: Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$



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

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.

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:

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