

Machine Learning

Max-Likelihood Principle, Naïve Bayes Model

DSC 240

Feb 13, 2025

Instructor: Prof. Yu-Xiang Wang

MP3: Predict Credit Card Approval

- Still binary classification but more open ended
 - you can use off-the-shelf tool “sklearn” for feature processing, defining and training classifiers.
- You will compete in a leaderboard. Top 10 students (in the final test result) get a bonus.
- During competition, leaderboard shows your result on public test set (your “dev” set). After the deadline, we will reveal results on a clean private data.

Last lecture

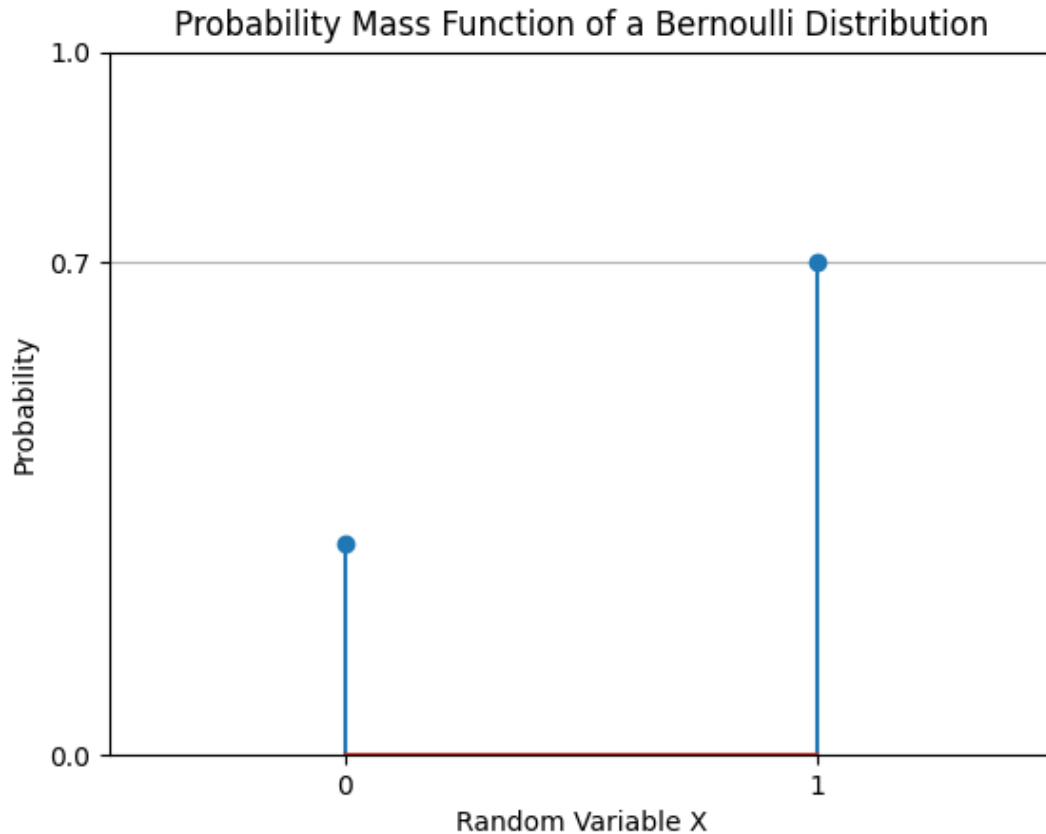
- Max-margin linear separator --- hinge loss + L2 regularization
 - Also known as the notorious “support vector machines”
- Probability and statistics review
 - A fundamental problem of statistics: estimation!
 - How to estimate quantities about the population using sampled data?
 - Example: Polls.
 - Example: Estimating a biased coin. Estimating average precipitation.

Today

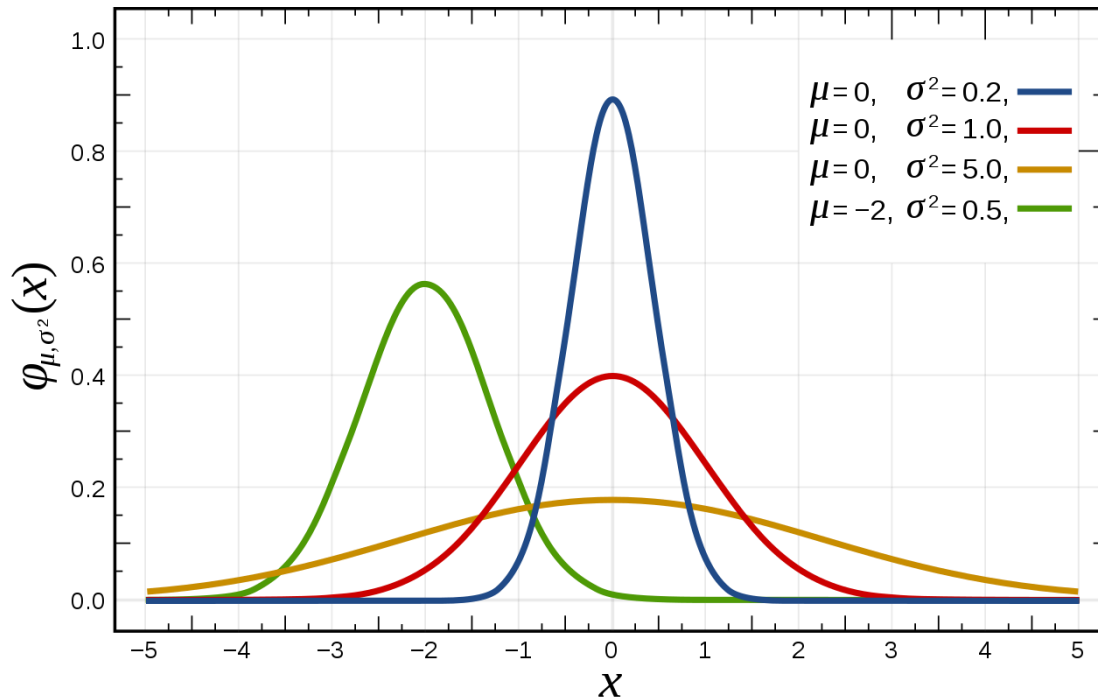
- Probabilistic models and max-likelihood estimation
- Derive square loss and logistic loss
- Naïve Bayes models

Bernoulli Distribution $X \sim \text{Ber}(p)$

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



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



(figure from wikipedia)

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

Maximum likelihood estimation

- Used since Gauss, Laplace, etc.... Popularized / carefully analyzed by Ronald Fisher.
- Which distribution is more *likely* to have produced the data?

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

- Observation 1: If the data is i.i.d. the by independence the density factorizes
- Observation 2: Taking log does not change the solution.

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

- $P(\text{Data}; \text{Parameter})$
 - 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.

What do we know about MLE? I won't prove these facts for you but... it's useful to know them.

- The error in parametric estimation goes to 0

$$\mathbb{E}\|\hat{\theta}_{\text{MLE}} - \theta^*\|_2 = O(1/\sqrt{n})$$

- MLE is equivalent to minimizing the KL-divergence

$$\min_{P \in \Pi} D_{KL}(P^*(\text{Data}) \| P(\text{Data}))$$

- It is asymptotically (as $n \rightarrow \infty$) the **most sample-efficient** estimator (in almost every statistical estimation problem...)

Exercise: Estimating Bernoulli Mean using MLE

- Data: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Bernoulli}(p)$
- Likelihood: $\mathcal{L}(X_i; p) = p^{X_i} (1 - p)^{1 - X_i}$
- The MLE problem: $\hat{p} = \arg \max_{p \in [0, 1]} p^{\sum_i X_i} (1 - p)^{n - \sum_i X_i}$

Exercise: Estimating the mean parameter of a Gaussian distribution

• Data $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$

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

Exercise: 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: $(x_1, y_1), \dots, (x_n, y_n)$
- Work out the optimization problem to solve for the MLE for θ^* .

Exercise: Logistic regression

- $P(y|x)$ is modeled by a Logit model $y \sim \text{Bernoulli}(\text{Sigmoid}(x^T \theta^*))$

$$\text{where } \text{Sigmoid}(t) = \frac{e^t}{e^t + 1}$$

- Data: $(x_1, y_1), \dots, (x_n, y_n)$
- 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

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

- For regression problems

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

Checkpoint

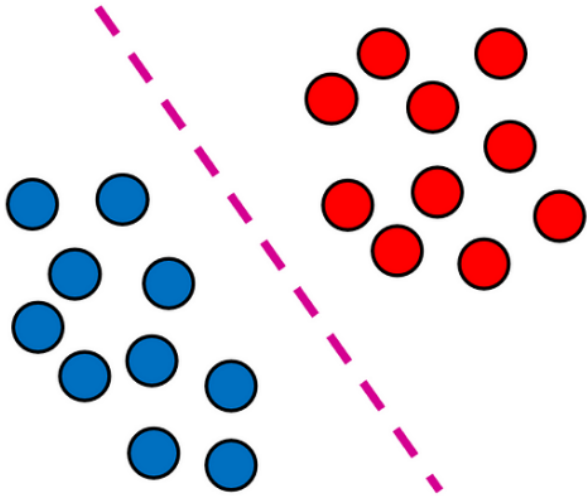
- We have shown that
 - Square loss minimization is MLE under a Gaussian noise model
 - Logistic loss minimization is MLE under a Bernoulli model for binary classification.

Recap: 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***.
- The natural prediction would be
 - $h(x) = \operatorname{argmax}_y P(y | x)$
 - If the data generative process is indeed $P(y|x)$, then this is “Bayes optimal”.

Discriminative models vs generative models

Discriminative



Generative

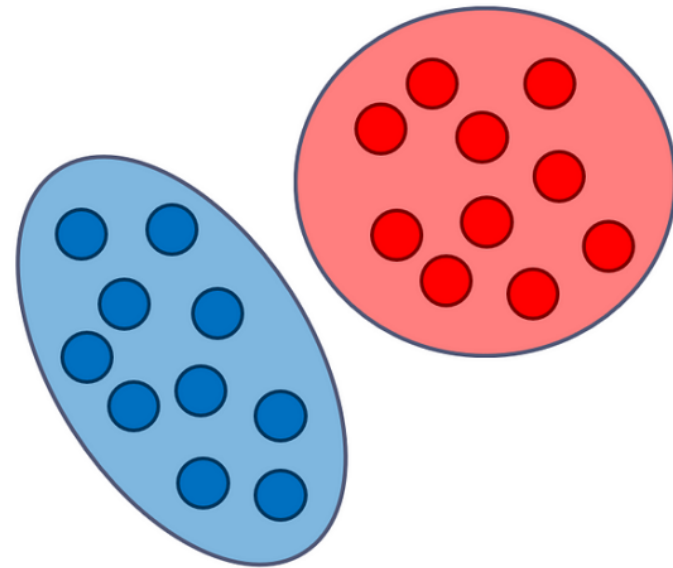


Image Credit: Dr. Roi Yehoshua

Generative model builds a world. How do we model the joint distribution $p(x,y)$?

- Prior: $p(y)$
- Per-class generative distribution $p(x|y)$
- Then how do we make inference about the label given feature?
 - By Bayes rule
- How to we determine $p(y)$ and $p(x|y)$?
 - Fit them using data by MLE!

Modeling $p(x|y)$ is challenging

Consider a dataset with 16 attributes (lets assume they are all binary).

How many parameters to we need to estimate to fully determine $p(\mathbf{X}|Y)$?

age	employe	education	edun	marital	...	job	relation	race	gender	hour	country	wealth
39	State_gov	Bachelors	13	Never_mar...	...	Adm_cleric	Not_in_fan	White	Male	40	United_Stat	poor
51	Self_emp	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Stat	poor
39	Private	HS_grad	9	Divorced	...	Handlers_c	Not_in_fan	White	Male	40	United_Stat	poor
54	Private	11th	7	Married	...	Handlers_c	Husband	Black	Male	40	United_Stat	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Stat	poor
50	Private	9th	5	Married_sp...	...	Other_serv	Not_in_fan	Black	Female	16	Jamaica	poor
52	Self_emp	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Stat	rich
31	Private	Masters	14	Never_mar...	...	Prof_speci	Not_in_fan	White	Female	50	United_Stat	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Stat	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Stat	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar...	...	Adm_cleric	Own_child	White	Female	30	United_Stat	poor
33	Private	Assoc_acc	12	Never_mar...	...	Sales	Not_in_fan	Black	Male	50	United_Stat	poor
41	Private	Assoc_voc	11	Married	...	Craft_repai	Husband	Asian	Male	40	*MissingV	rich
34	Private	7th_8th	4	Married	...	Transport	Husband	Amer_Indic	Male	45	Mexico	poor
26	Self_emp	HS_grad	9	Never_mar...	...	Farming_fi	Own_child	White	Male	35	United_Stat	poor
33	Private	HS_grad	9	Never_mar...	...	Machine_c	Unmarried	White	Male	40	United_Stat	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Stat	poor
44	Self_emp	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Stat	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Stat	rich

Learning the values for the full conditional probability table would require enormous amounts of data.

Simple example

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 parameters required to determine $p(X|y)$?

2. What happens if there are d Boolean features?

Naïve Bayes assumption

- Naïve Bayes classifiers assume that given the class label Y the features are **conditional independent** of each other

$$p(\mathbf{X}|y) = \prod_j p_j(x^j|Y)$$

- p_j : specific model for attribute j .

Simple example with naïve Bayes assumption

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 parameters required to determine $p(X|y)$?

2. What happens if there are d Boolean features?

Example: Text classification

Machine learning is a subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. In other words, it's a process of data analysis that automates analytical model building. Machine learning involves the creation and use of algorithms that can learn from and make decisions or predictions based on data...



Human ?



Machine ?

X is a sequence of words

y is "Human" or "Machine"

Naïve Bayes Assumption:

$P(X | \text{"Machine"}) = P(\text{Word 1} | \text{"Machine"}) \dots P(\text{Word N} | \text{"machine"})$

MLE for fitting the Naïve Bayes Model for text classification

All machine generated texts.
(m documents
 i th document with N_i
words.)

$$\hat{p}(y = \text{"machine"}) = \frac{m}{m + n}$$

$$\hat{p}(\text{Word} = w | y = \text{"machine"}) = \frac{\# \text{ of } w}{\sum_{i=1}^m N_i}$$

All human written text.
(n documents. j th
document with N_j words.)

$$\hat{p}(y = \text{"human"}) = \frac{n}{m + n}$$

$$\hat{p}(\text{Word} = w | y = \text{"human"}) = \frac{\# \text{ of } w}{\sum_{j=1}^n N_j}$$

2 min exercise: Naïve Bayes Model for text classification

Sentence 1: "Why do you cry?"
Sentence 2: "Yes."
Sentence 3: "Hasta La Vista,
Baby!"
Sentence 4: "I will be back."

Sentence 1: "You mean
people?"
Sentence 2 "I don't know. We
just cry."
Sentence 3: "No, no, no, no.
You gotta listen to the way
people talk. "

- What is our estimate of $P(y)$?
- What is our estimate for $P(\text{word} = \text{"No"} | y)$ and $P(\text{word} = \text{"You"} | y)$?

Prediction with Naïve Bayes Model: “Just plug in”

Once we computed all parameters for attributes in both classes we can easily decide on the label of a **new** sample \mathbf{x} :

$$\begin{aligned}\hat{y} &= \operatorname{argmax}_{y_k} p(y = y_k | \mathbf{x}) \\ &= \operatorname{argmax}_{y_k} \frac{p(\mathbf{x} | y = y_k) p(y = y_k)}{p(\mathbf{x})} \\ &= \operatorname{argmax}_{y_k} \prod_j p(\mathbf{x}^j | y = y_k) p(y = y_k)\end{aligned}$$

Try it on this example:
New sentence: “I know now why you cry”

- Perform this computation for both class 1 and class 2 and select the class that leads to a higher probability as your decision

Naïve Bayes model with continuous variables

- So far we assumed a binomial or discrete distribution for the data given the model ($p(\mathbf{x}^i|y)$)
- However, in many cases the data contains continuous features:
 - Height, weight, Levels of genes in cells, Brain activity
- **Gaussian Naïve Bayes model:**

$$x_i|y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

You will learn more about Gaussian Naïve Bayes models in HW3!

Next Lecture

- Error decomposition
- Decision Tree and Boosting
- Gradient Boosting