UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Fall 2024)

Machine Learning

Lecture 13: Naïve Bayes Models

Chong Liu

Assistant Professor of Computer Science

Oct 10, 2024

Today

- Maximum likelihood estimation
  - Linear regression
- Naïve Bayes models
- Midterm review
  - What you should know

Recap: Estimating mean of Gaussian distr.

- Data

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

- 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:

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

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:

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

- 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

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

Prediction after 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:

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

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

Recap: directing modeling $p(x|y)$ is hard

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. At least 8 data points are needed to determine $p(x|y)$.

2. At least 2^d data points are needed if there are d binary features.

Naïve Bayes assumption

- Given the class label y, the features are conditional independent of each other.
  - $P(x|y) = \prod_j p(x_j|y)$
  - x_j is the j-th feature, not j-th data point!

- How to understand?
  - All features independently affect the label.

Model $p(x|y)$ 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

Discussion: How many data points are needed to determine $p(x|y)$ given binary output and d binary features?

$$2d$$

Why? When we use the first feature, we ignore all other features…

Naïve Bayes model

- Goal of probabilistic model:
  - $h(x) = \text{argmax}_y P(y|x)$
  - $h(x) = \text{argmax}_y \frac{P(y)P(x|y)}{P(x)}$
  - $h(x) = \text{argmax}_y P(y)P(x|y)$
  - $h_{\text{Naïve\_Bayes}}(x) = \text{argmax}_y P(y) \prod_{j=1}^{d} P(x_j|y)$
- Bayes rule
  - $P(y|x) = \frac{P(y)P(x|y)}{P(x)}$
- Naïve Bayes assumption
  - $P(x|y) = \prod_j p(x_j|y)$

Example: Naïve Bayes model for email filter

Email ID	Contains "buy"	Contains "cheap"	Contains "free"	Label
1	Yes	Yes	No	Spam
2	No	No	Yes	Not Spam
3	Yes	No	Yes	Spam
4	No	Yes	No	Not Spam
5	Yes	Yes	Yes	Spam
6	No	No	No	Not Spam

- Naïve Bayes model: $h_{\text{Naïve\_Bayes}}(x) = \text{argmax}_y P(y) \prod_{j=1}^{d} P(x_j|y)$
- Step 1: calculate $P(y)$
- Step 2: calculate $P(x_j|y)$
- Step 3: prediction on [Yes, No, No]

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:**

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

Checkpoint: Probabilistic models

- Probabilistic / Generative / Bayesian models are very powerful and interpretable method for modeling the world.
  - Customize ML models for your applications.
  - Explicitly model the dependence.
  - Naïve Bayes model is the simplest of them all!

Midterm exam

- What does the exam look like?
  - 70 min (12:05-1:15pm) on Thu Oct 24 at LC 5
  - Please arrive at 12pm!
  - Closed-book exam
  - Given individually (not in groups!)
  - Counts 20% towards your final grades
  - No make-up exam
- What to bring?
  - Your pen
- What not to bring?
  - Your book, note, lecture slide, or cheat sheet.

What are you expected to know?

- Basic mathematical tools
  - In our math review (Lecture 2-4)
    - Linear algebra, calculus and optimization, probability and statistics
  - Review Homework 1!

What are you expected to know?

- Basic concepts of machine learning
  - Classification and regression
  - Input space (feature space), output space (label space), hypothesis class
  - Confusion matrix of binary classification
  - Accuracy
  - Holdout / cross validation / hyperparameter
  - Problem of overfitting
  - Loss function
  - Linear model

What are you expected to know?

- Understanding how machine learning algorithms work
  - Why do we need surrogate loss in classification?
  - Why do we need SGD? Drawback of GD?
  - How to define a linear classifier / linear regression?
  - Why do we need SVM? Difference between linear classifier and SVM.
  - Why do we need regularization? How to apply it?
  - Key idea of maximum likelihood estimation.
  - Key assumption of Naïve Bayes models.

Announcement

- Homework 1 grades have been released.
  - Ask your group member who submitted the solution
- Instructor office hours
  - Cancelled on Tue Oct 15 & Tue Oct 22
  - A make-up office hour is scheduled at
    - 2-3pm on Thu Oct 17 at UAB 426 (last one before midterm exam!)
- Midterm presentation (Thu Oct 17)
  - 15 groups, each has 4 min
  - No credit towards your final grades
  - Use it as a practice opportunity!
  - Send slides to me by 11:59pm Wed Oct 16