

Probabilistic Modeling

He He

CDS, NYU

March 24, 2020

Contents

1 Overview

2 Conditional models

- Recap: linear regression
- A probabilistic view of linear regression
- Logistic regression
- Generalized Linear Models

3 Generative models

- Naive Bayes models
 - Bernoulli Naive Bayes Models
 - Gaussian Naive Bayes Models

- Staff change for the second half
 - **Instructor:** He He
 - Office hours: Tue 4:00-5:00pm (before the lecture), Zoom / CDS 605
 - **Section lead:** Shubam Chandel
 - Office hours: Wed 5:30-6:30pm (before the section), Zoom / CDS 650
- Course material
 - No required readings but resources will be posted by the Friday before the lecture
- Submissions
 - Homework 4 due this Friday
 - Project proposal due this Friday
 - “Homework”: project_0 / Codalab
- Please complete the teaching evaluation on [Quadratics](#)

Overview

Why probabilistic modeling?

- A unified framework that covers many models, e.g., linear regression, logistic regression
- Learning as **statistical inference**
- Principled ways to incorporate your belief on the data generating distribution (inductive biases)

Today's lecture

- Two ways to model how the data is generated:

Today's lecture

- Two ways to model how the data is generated:
 - **Conditional:** $p(y \mid x)$
 - **Generative:** $p(x, y)$

Today's lecture

- Two ways to model how the data is generated:
 - **Conditional:** $p(y | x)$
 - **Generative:** $p(x, y)$
- How to estimate the parameters of our model? Maximum likelihood estimation.

Today's lecture

- Two ways to model how the data is generated:
 - **Conditional:** $p(y | x)$
 - **Generative:** $p(x, y)$
- How to estimate the parameters of our model? Maximum likelihood estimation.
- Compare and contrast conditional and generative models.

Conditional models

Linear regression

Linear regression is one of the most important methods in machine learning and statistics.

Goal: Predict a real-valued **target** y (also called response) from a vector of **features** x (also called covariates).

Linear regression

Linear regression is one of the most important methods in machine learning and statistics.

Goal: Predict a real-valued **target** y (also called response) from a vector of **features** x (also called covariates).

Examples:

- Predicting house price given location, condition, build year etc.
- Predicting medical cost of a person given age, sex, region, BMI etc.
- Predicting age of a person based on their photos.

Problem setup

Data Training examples $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

Problem setup

Data Training examples $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

Model A *linear* function h (parametrized by θ) to predict y from x :

$$h(x) = \sum_{i=0}^d \theta_i x_i = \theta^T x, \quad (1)$$

where $\theta \in \mathbb{R}^d$ are the **parameters** (also called weights).

Problem setup

Data Training examples $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

Model A *linear* function h (parametrized by θ) to predict y from x :

$$h(x) = \sum_{i=0}^d \theta_i x_i = \theta^T x, \quad (1)$$

where $\theta \in \mathbb{R}^d$ are the **parameters** (also called weights).

Note that

- We incorporate the **bias term** (also called the intercept term) into x (i.e. $x_0 = 1$).
- We use superscript to denote the example id and subscript to denote the dimension id.

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2. \quad (\text{empirical risk}) \quad (2)$$

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2. \quad (\text{empirical risk}) \quad (2)$$

- Matrix form**
- Let $X \in \mathbb{R}^{N \times d}$ be the **design matrix** whose rows are input features.
 - Let $\mathbf{y} \in \mathbb{R}^N$ be the vector of all targets.

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2. \quad (\text{empirical risk}) \quad (2)$$

- Matrix form**
- Let $X \in \mathbb{R}^{N \times d}$ be the **design matrix** whose rows are input features.
 - Let $\mathbf{y} \in \mathbb{R}^N$ be the vector of all targets.
 - We want to solve

$$\hat{\theta} = \arg \min_{\theta} (X\theta - \mathbf{y})^T (X\theta - \mathbf{y}). \quad (3)$$

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2. \quad (\text{empirical risk}) \quad (2)$$

- Matrix form**
- Let $X \in \mathbb{R}^{N \times d}$ be the **design matrix** whose rows are input features.
 - Let $\mathbf{y} \in \mathbb{R}^N$ be the vector of all targets.
 - We want to solve

$$\hat{\theta} = \arg \min_{\theta} (X\theta - \mathbf{y})^T (X\theta - \mathbf{y}). \quad (3)$$

Solution Closed-form solution: $\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}$.

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2. \quad (\text{empirical risk}) \quad (2)$$

- Matrix form**
- Let $X \in \mathbb{R}^{N \times d}$ be the **design matrix** whose rows are input features.
 - Let $\mathbf{y} \in \mathbb{R}^N$ be the vector of all targets.
 - We want to solve

$$\hat{\theta} = \arg \min_{\theta} (X\theta - \mathbf{y})^T (X\theta - \mathbf{y}). \quad (3)$$

Solution Closed-form solution: $\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}$.

Review questions

- Derive the solution for linear regression.
- What if $X^T X$ is not invertible?

We've seen

- Linear regression: response is a linear function of the inputs
- Estimate parameters by minimize the squared loss

We've seen

- Linear regression: response is a linear function of the inputs
- Estimate parameters by minimize the squared loss

But...

- Why squared loss is a reasonable choice for regression problems?
- What assumptions are we making on the data? (**inductive bias**)

We've seen

- Linear regression: response is a linear function of the inputs
- Estimate parameters by minimize the squared loss

But...

- Why squared loss is a reasonable choice for regression problems?
- What assumptions are we making on the data? (inductive bias)

Next,

- Derive linear regression from a probabilistic modeling perspective.

Assumptions in linear regression

Assumptions in linear regression

- $x^{(n)}$ and $y^{(n)}$ are related through a linear function:

$$y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}, \quad (4)$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

Assumptions in linear regression

- $x^{(n)}$ and $y^{(n)}$ are related through a linear function:

$$y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}, \quad (4)$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

- The errors are distributed *iid* (independently and identically distributed):

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \quad (5)$$

Assumptions in linear regression

- $x^{(n)}$ and $y^{(n)}$ are related through a linear function:

$$y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}, \quad (4)$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

- The errors are distributed *iid* (independently and identically distributed):

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \quad (5)$$

What's the distribution of $Y | X$?

Assumptions in linear regression

- $x^{(n)}$ and $y^{(n)}$ are related through a linear function:

$$y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}, \quad (4)$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

- The errors are distributed *iid* (independently and identically distributed):

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \quad (5)$$

What's the distribution of $Y | X$?

$$p(y | x; \theta) = \mathcal{N}(\theta^T x, \sigma^2). \quad (6)$$

Imagine putting a Gaussian bump around the output of the linear predictor.

Assumptions in linear regression

- $x^{(n)}$ and $y^{(n)}$ are related through a linear function:

$$y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}, \quad (4)$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

- The errors are distributed *iid* (independently and identically distributed):

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \quad (5)$$

What's the distribution of $Y | X$?

$$p(y | x; \theta) = \mathcal{N}(\theta^T x, \sigma^2). \quad (6)$$

θ as a fixed parameter vs a variable (more on this next time!).

Imagine putting a Gaussian bump around the output of the linear predictor.

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset \mathcal{D} , how to estimate the model parameters θ ?

(8)

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset \mathcal{D} , how to estimate the model parameters θ ?

The **maximum likelihood principle** says that we should maximize the (conditional) likelihood of the data:

$$L(\theta) \stackrel{\text{def}}{=} p(\mathcal{D}; \theta) \tag{7}$$

(8)

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset \mathcal{D} , how to estimate the model parameters θ ?

The **maximum likelihood principle** says that we should maximize the (conditional) likelihood of the data:

$$L(\theta) \stackrel{\text{def}}{=} p(\mathcal{D}; \theta) \tag{7}$$

$$= \prod_{n=1}^N p(y^{(n)} \mid x^{(n)}; \theta). \tag{8}$$

(examples are distributed *iid*)

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset \mathcal{D} , how to estimate the model parameters θ ?

The **maximum likelihood principle** says that we should maximize the (conditional) likelihood of the data:

$$L(\theta) \stackrel{\text{def}}{=} p(\mathcal{D}; \theta) \tag{7}$$

$$= \prod_{n=1}^N p(y^{(n)} \mid x^{(n)}; \theta). \tag{8}$$

(examples are distributed *iid*)

In practice, we maximize the **log likelihood** $\ell(\theta)$, or equivalently, minimize the negative log likelihood (NLL).

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset \mathcal{D} , how to estimate the model parameters θ ?


The **maximum likelihood principle** says that we should maximize the (conditional) likelihood of the data:

$$L(\theta) \stackrel{\text{def}}{=} p(\mathcal{D}; \theta) \tag{7}$$

$$= \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta). \tag{8}$$

(examples are distributed *iid*)

In practice, we maximize the **log likelihood** $\ell(\theta)$, or equivalently, minimize the negative log likelihood (NLL).


$$L(\theta_1) \leq L(\theta_2) \implies \log L(\theta_1) \leq \log L(\theta_2)$$

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

(13)

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta) \tag{9}$$

$$= \log \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta) \tag{10}$$

$$= \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \tag{11}$$

(13)

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta) \tag{9}$$

$$= \log \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta) \tag{10}$$

$$= \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \tag{11}$$

$$= \sum_{n=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(n)} - \theta^T x^{(n)})^2}{2\sigma^2}\right) \tag{12}$$

$$\tag{13}$$

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta) \tag{9}$$

$$= \log \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta) \tag{10}$$

$$= \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \tag{11}$$

$$= \sum_{n=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(n)} - \theta^T x^{(n)})^2}{2\sigma^2}\right) \tag{12}$$

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N (y^{(n)} - \theta^T x^{(n)})^2 \tag{13}$$

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta) \tag{9}$$

$$= \log \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta) \tag{10}$$

$$= \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \tag{11}$$

$$= \sum_{n=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(n)} - \theta^T x^{(n)})^2}{2\sigma^2}\right) \tag{12}$$

does not depend on θ

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N (y^{(n)} - \theta^T x^{(n)})^2 \tag{13}$$

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y | X \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta) \tag{9}$$

$$= \log \prod_{n=1}^N p(y^{(n)} | x^{(n)}; \theta) \tag{10}$$

$$= \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \tag{11}$$

$$= \sum_{n=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y^{(n)} - \theta^T x^{(n)})^2}{2\sigma^2} \right)$$

does not depend on θ

minimizing squared loss!

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N (y^{(n)} - \theta^T x^{(n)})^2 \tag{13}$$

Gradient of the likelihood

Recall that we obtained the normal equation by setting the derivative of the squared loss to zero. Now let's compute the derivative of the likelihood w.r.t. the parameters.

$$\ell(\theta) = N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2 \quad (14)$$

(15)

Gradient of the likelihood

Recall that we obtained the normal equation by setting the derivative of the squared loss to zero. Now let's compute the derivative of the likelihood w.r.t. the parameters.

$$\ell(\theta) = N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2 \quad (14)$$

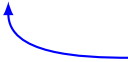
$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{\sigma^2} \sum_{n=1}^N (y^{(n)} - \theta^T x^{(n)}) x_i^{(n)}. \quad (15)$$

Gradient of the likelihood

Recall that we obtained the normal equation by setting the derivative of the squared loss to zero. Now let's compute the derivative of the likelihood w.r.t. the parameters.

$$\ell(\theta) = N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^N \left(y^{(n)} - \theta^T x^{(n)} \right)^2 \quad (14)$$

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{\sigma^2} \sum_{n=1}^N (y^{(n)} - \theta^T x^{(n)}) x_i^{(n)}. \quad (15)$$



$\mathbb{E}[Y | X = x^{(n)}]$

(Spoiler: we will see this form again.)

We've seen

- Linear regression assumes that $Y | X$ follows a Gaussian distribution
- MLE of linear regression is equivalent to the least square method

We've seen

- Linear regression assumes that $Y | X$ follows a Gaussian distribution
- MLE of linear regression is equivalent to the least square method

However,

- Sometimes Gaussian distribution is not a reasonable assumption, e.g., classification
- Can we use the same modeling approach for other prediction tasks?

We've seen

- Linear regression assumes that $Y | X$ follows a Gaussian distribution
- MLE of linear regression is equivalent to the least square method

However,

- Sometimes Gaussian distribution is not a reasonable assumption, e.g., classification
- Can we use the same modeling approach for other prediction tasks?

Next,

- Derive [logistic regression](#) for classification.

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

How should we parameterize $h(x)$?

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

How should we parameterize $h(x)$?

- What is $p(y = 1 | x)$ and $p(y = 0 | x)$?

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

How should we parameterize $h(x)$?

- What is $p(y = 1 | x)$ and $p(y = 0 | x)$? $h(x) \in (0, 1)$.

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

How should we parameterize $h(x)$?

- What is $p(y = 1 | x)$ and $p(y = 0 | x)$? $h(x) \in (0, 1)$.
- What is the mean of $Y | X$?

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

How should we parameterize $h(x)$?

- What is $p(y = 1 | x)$ and $p(y = 0 | x)$? $h(x) \in (0, 1)$.
- What is the mean of $Y | X$? $h(x)$. (Think how we parameterize the mean in linear regression)

Assumptions in logistic regression

Consider binary classification where $Y \in \{0, 1\}$. What distribution could $Y | X$ follow?

We model $p(y | x)$ as a **Bernoulli** distribution:

$$p(y | x) = h(x)^y (1 - h(x))^{1-y}. \quad (16)$$

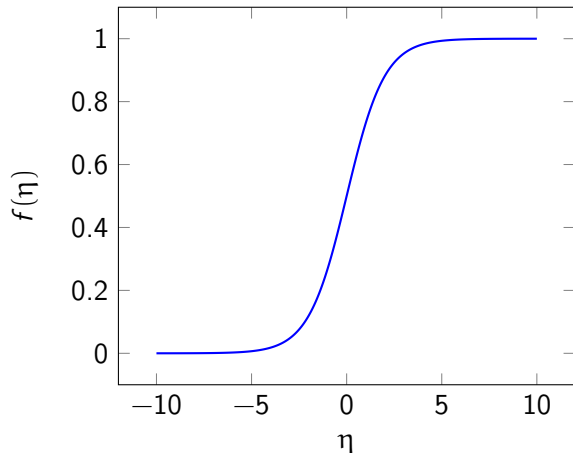
How should we parameterize $h(x)$?

- What is $p(y = 1 | x)$ and $p(y = 0 | x)$? $h(x) \in (0, 1)$.
- What is the mean of $Y | X$? $h(x)$. (Think how we parameterize the mean in linear regression)
- Need a function f to map the linear predictor $\theta^T x$ in \mathbb{R} to $(0, 1)$:

$$f(\eta) = \frac{1}{1 + e^{-\eta}} \quad \text{logistic function} \quad (17)$$

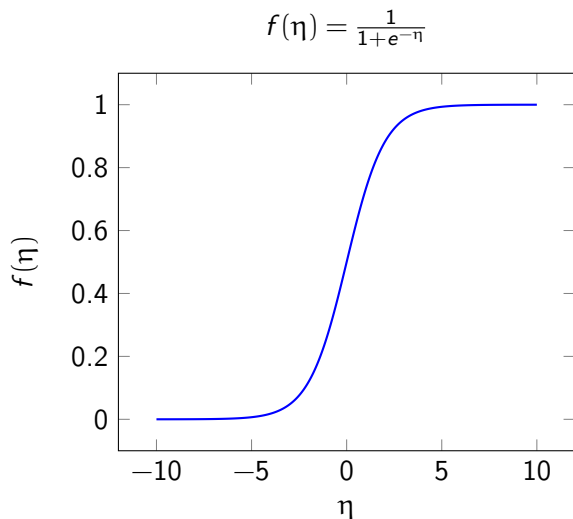
Logistic regression

$$f(\eta) = \frac{1}{1+e^{-\eta}}$$



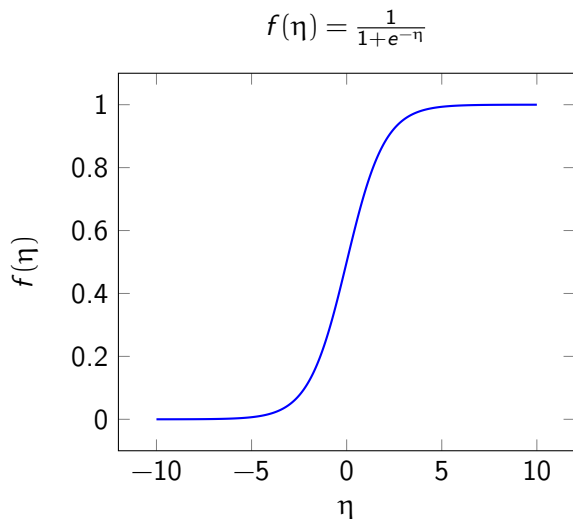
- $p(y | x) = \text{Bernoulli}(f(\theta^T x)).$

Logistic regression



- $p(y | x) = \text{Bernoulli}(f(\theta^T x))$.
- When do we have $p(y = 1 | x) = 1$ and $p(y = 0 | x) = 1$?

Logistic regression

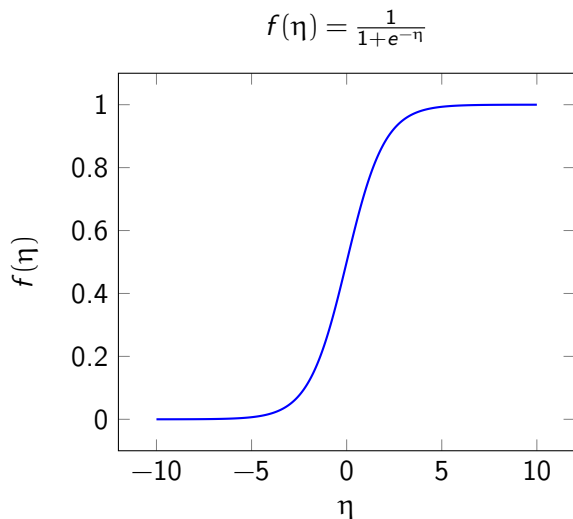


- $p(y | x) = \text{Bernoulli}(f(\theta^T x))$.
- When do we have $p(y = 1 | x) = 1$ and $p(y = 0 | x) = 1$?
- **Exercise:** show that the **log odds** is

$$\log \frac{p(y = 1 | x)}{p(y = 0 | x)} = \theta^T x. \quad (18)$$

$$\implies \text{linear decision boundary} \quad (19)$$

Logistic regression



- $p(y | x) = \text{Bernoulli}(f(\theta^T x))$.
- When do we have $p(y = 1 | x) = 1$ and $p(y = 0 | x) = 1$?
- **Exercise:** show that the **log odds** is

$$\log \frac{p(y = 1 | x)}{p(y = 0 | x)} = \theta^T x. \quad (18)$$

$$\implies \text{linear decision boundary} \quad (19)$$

- How do we extend it to multiclass classification? (more on this later)

MLE for logistic regression

Similar to linear regression, let's estimate θ by maximizing the conditional log likelihood.

(21)

MLE for logistic regression

Similar to linear regression, let's estimate θ by maximizing the conditional log likelihood.

$$\ell(\theta) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \quad (20)$$

(21)

MLE for logistic regression

Similar to linear regression, let's estimate θ by maximizing the conditional log likelihood.

$$\ell(\theta) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \quad (20)$$

$$= \sum_{n=1}^N y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)})) \quad (21)$$

MLE for logistic regression

Similar to linear regression, let's estimate θ by maximizing the conditional log likelihood.

$$\ell(\theta) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}; \theta) \quad (20)$$

$$= \sum_{n=1}^N y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)})) \quad (21)$$

- Closed-form solutions are not available.
- But, the likelihood is convex — [gradient ascent](#) gives us the unique optimal solution.

$$\theta := \theta + \alpha \nabla_{\theta} \ell(\theta). \quad (22)$$

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

(26)

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

(26)

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

$$\frac{\partial \ell^n}{\partial \theta_i} = \frac{\partial \ell^n}{\partial f^n} \frac{\partial f^n}{\partial \theta_i} \quad (23)$$

(26)

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

$$\frac{\partial \ell^n}{\partial \theta_i} = \frac{\partial \ell^n}{\partial f^n} \frac{\partial f^n}{\partial \theta_i} \quad (23)$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1 - y^{(n)}}{1 - f^n} \right) \frac{\partial f^n}{\partial \theta_i} \quad \frac{d}{dx} \ln x = \frac{1}{x} \quad (24)$$

(26)

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

$$\frac{\partial \ell^n}{\partial \theta_i} = \frac{\partial \ell^n}{\partial f^n} \frac{\partial f^n}{\partial \theta_i} \quad (23)$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1 - y^{(n)}}{1 - f^n} \right) \frac{\partial f^n}{\partial \theta_i} \quad \frac{d}{dx} \ln x = \frac{1}{x} \quad (24)$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1 - y^{(n)}}{1 - f^n} \right) \left(f^n(1 - f^n) x_i^{(n)} \right) \quad \text{Exercise: apply chain rule to } \frac{\partial f^n}{\partial \theta_i} \quad (25)$$

$$= (y^{(n)} - f^n) x_i^{(n)} \quad \text{simplify by algebra} \quad (26)$$

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x , e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

$$\frac{\partial \ell^n}{\partial \theta_i} = \frac{\partial \ell^n}{\partial f^n} \frac{\partial f^n}{\partial \theta_i} \quad (23)$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1 - y^{(n)}}{1 - f^n} \right) \frac{\partial f^n}{\partial \theta_i} \quad \frac{d}{dx} \ln x = \frac{1}{x} \quad (24)$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1 - y^{(n)}}{1 - f^n} \right) \left(f^n (1 - f^n) x_i^{(n)} \right) \quad \text{Exercise: apply chain rule to } \frac{\partial f^n}{\partial \theta_i} \quad (25)$$

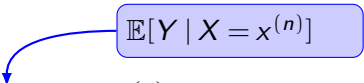
$$= (y^{(n)} - f^n) x_i^{(n)} \quad \text{simplify by algebra} \quad (26)$$

The full gradient is thus $\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^N (y^{(n)} - f(\theta^T x^{(n)})) x_i^{(n)}$.

A closer look at the gradient

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^N (y^{(n)} - f(\theta^T x^{(n)})) x_i^{(n)} \quad (27)$$


A closer look at the gradient

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^N (y^{(n)} - \textcolor{blue}{f(\theta^T x^{(n)})}) x_i^{(n)} \quad (27)$$


A blue curved arrow points from the boxed expression $\mathbb{E}[Y | X = x^{(n)}]$ to the function f in the equation above. The function f and the term $x^{(n)}$ in the argument are also colored blue.

A closer look at the gradient

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^N (y^{(n)} - f(\theta^T x^{(n)})) x_i^{(n)} \quad (27)$$

 $\mathbb{E}[Y | X = x^{(n)}]$

- Does this look familiar?
- Our derivation for linear regression and logistic regression are quite similar...
- Next, a more general family of models.

Compare linear regression and logistic regression

	linear regression	logistic regression
--	-------------------	---------------------

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)
Response type	real	categorical
Response distribution	Gaussian	Bernoulli

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)
Response type	real	categorical
Response distribution	Gaussian	Bernoulli
Response function $f(\theta^T x)$	identity	logistic

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)
Response type	real	categorical
Response distribution	Gaussian	Bernoulli
Response function $f(\theta^T x)$	identity	logistic
Response mean $\mathbb{E}(Y X = x; \theta)$	$f(\theta^T x)$	$f(\theta^T x)$

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)
Response type	real	categorical
Response distribution	Gaussian	Bernoulli
Response function $f(\theta^T x)$	identity	logistic
Response mean $\mathbb{E}(Y X = x; \theta)$	$f(\theta^T x)$	$f(\theta^T x)$

- x enters through a linear function.
- The main difference between the formulations is due to different response variables.
- Can we generalize the idea to handle other response types, e.g., positive integers?

Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:

$$p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta)) \quad (28)$$


Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:

$$p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta)) \quad (28)$$

natural parameters



Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:

$$p(y; \eta) = h(y) \exp(\underbrace{\eta^T}_{\text{natural parameters}} \underbrace{T(y)}_{\text{sufficient statistics}} - A(\eta)) \quad (28)$$

Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:

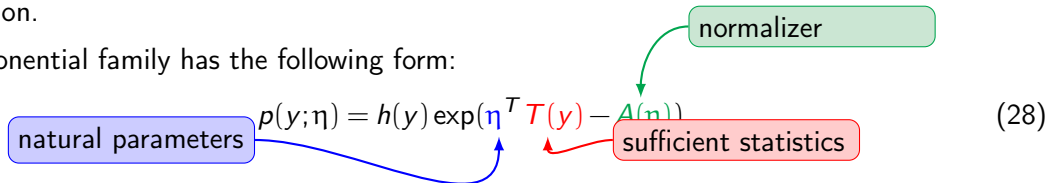
$$p(y; \eta) = h(y) \exp(\underbrace{\eta^T}_{\text{natural parameters}} \underbrace{T(y)}_{\text{sufficient statistics}} - \underbrace{A(\eta)}_{\text{normalizer}}) \quad (28)$$

The diagram illustrates the components of the exponential family distribution formula. A blue box labeled "natural parameters" points to the η term. A red box labeled "sufficient statistics" points to the $T(y)$ term. A green box labeled "normalizer" points to the $A(\eta)$ term.

Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:



The diagram shows the probability mass function of the exponential family: $p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta))$. Annotations include: a blue box labeled 'natural parameters' with an arrow pointing to η ; a red box labeled 'sufficient statistics' with an arrow pointing to $T(y)$; and a green box labeled 'normalizer' with an arrow pointing to $A(\eta)$. The terms η , $T(y)$, and $A(\eta)$ are color-coded to match their respective boxes (blue, red, and green).

$$p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta)) \quad (28)$$

- Many nice properties, e.g., conjugate priors for Bayesian inference.
- **Useful property for this class:**

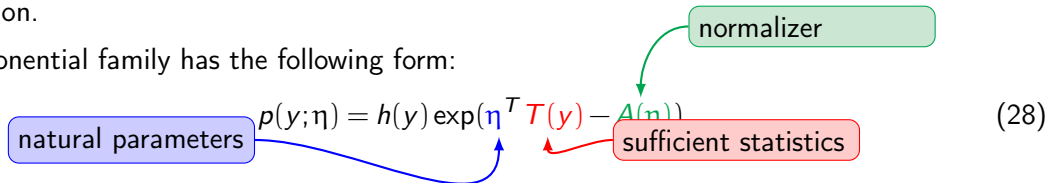
$$\mu \stackrel{\text{def}}{=} \mathbb{E}[Y] = \psi^{-1}(\eta), \quad \eta = \psi(\mu) \quad (29)$$

$$\mu = A'(\eta) \quad (30)$$

Generalized linear models (GLM)

Main idea: use a class of distributions known as the **exponential family** as the response distribution.

The exponential family has the following form:



The diagram shows the exponential family distribution formula: $p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta))$. Annotations include: a blue box labeled 'natural parameters' pointing to η ; a red box labeled 'sufficient statistics' pointing to $T(y)$; a green box labeled 'normalizer' pointing to $A(\eta)$. The equation is labeled (28) on the right.

$$p(y; \eta) = h(y) \exp(\eta^T T(y) - A(\eta)) \quad (28)$$

- Many nice properties, e.g., conjugate priors for Bayesian inference.
- **Useful property for this class:**

$$\mu \stackrel{\text{def}}{=} \mathbb{E}[Y] = \psi^{-1}(\eta), \quad \eta = \psi(\mu) \quad (29)$$

$$\mu = A'(\eta) \quad (30)$$

- Example: Gaussian, Bernoulli, Poisson distribution.
- **Exercise:** find η , $T(y)$, $A(\eta)$, $h(y)$ for Gaussian distribution.

Construct a GLM

Goal: build a model $h(x)$ to estimate $T(y)$ given some inputs x ($T(y) = y$ in most examples).

Assumption: an exponential family distribution is a good model for $Y | X$.

Construct a GLM

Goal: build a model $h(x)$ to estimate $T(y)$ given some inputs x ($T(y) = y$ in most examples).

Assumption: an exponential family distribution is a good model for $Y | X$.

- Choose a **response distribution**

$$Y | X = x; \theta \sim \text{ExponentialFamily}(\eta) \text{ where } \eta = \theta^T x. \quad (31)$$

Construct a GLM

Goal: build a model $h(x)$ to estimate $T(y)$ given some inputs x ($T(y) = y$ in most examples).

Assumption: an exponential family distribution is a good model for $Y | X$.

- Choose a **response distribution**

$$Y | X = x; \theta \sim \text{ExponentialFamily}(\eta) \text{ where } \eta = \theta^T x. \quad (31)$$

- Choose a **link function** f s.t.

$$f(\eta) = \mu = \mathbb{E}(Y | X = x). \quad (32)$$

- An available choice is ψ^{-1} , which is the canonical link function.

Construct a GLM

Goal: build a model $h(x)$ to estimate $T(y)$ given some inputs x ($T(y) = y$ in most examples).

Assumption: an exponential family distribution is a good model for $Y | X$.

- Choose a **response distribution**

$$Y | X = x; \theta \sim \text{ExponentialFamily}(\eta) \text{ where } \eta = \theta^T x. \quad (31)$$

- Choose a **link function** f s.t.

$$f(\eta) = \mu = \mathbb{E}(Y | X = x). \quad (32)$$

- An available choice is ψ^{-1} , which is the canonical link function.
- For prediction, use $h(x) = f(\eta) = \mathbb{E}(Y | X = x)$.

Example: Construct Poisson regression

Say we want to predict the number of people entering a restaurant in New York during lunch time.

- What features would be useful?
- What's a good model for number of visitors (the **response distribution**)?

Example: Construct Poisson regression

Say we want to predict the number of people entering a restaurant in New York during lunch time.

- What features would be useful?
- What's a good model for number of visitors (the **response distribution**)?

Math review: Poisson distribution

Given a random variable $Y \in 0, 1, 2, \dots$ following $\text{Poisson}(\lambda)$, we have


$$p(Y = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad (33)$$

where $\lambda > 0$ and $\mathbb{E}[Y] = \lambda$.

The Poisson distribution is usually used to model the number of events occurring during a fixed period of time.

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y \mid X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

(36)

(37)

(38)

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y \mid X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

(36)

(37)

(38)

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y | X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

$$p(y; \eta) = \exp \log \frac{\lambda^y e^{-\lambda}}{y!} \quad (34)$$

(36)

(37)

(38)

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y \mid X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

$$p(y; \eta) = \exp \log \frac{\lambda^y e^{-\lambda}}{y!} \quad (34)$$

$$= \exp(y \log \lambda - \lambda - \log y!) \quad (35)$$

$$(36)$$

$$(37)$$

$$(38)$$

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y | X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

$$p(y; \eta) = \exp \log \frac{\lambda^y e^{-\lambda}}{y!} \quad (34)$$

$$= \exp(y \log \lambda - \lambda - \log y!) \quad (35)$$

$$[\text{compare with } h(y) \exp(\eta^T T(y) - A(\eta))] \quad (36)$$

$$(37)$$

$$(38)$$

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y | X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

$$p(y; \eta) = \exp \log \frac{\lambda^y e^{-\lambda}}{y!} \quad (34)$$

$$= \exp(y \log \lambda - \lambda - \log y!) \quad (35)$$

$$[\text{compare with } h(y) \exp(\eta^T T(y) - A(\eta))] \quad (36)$$

$$(37)$$

$$\eta = \log \lambda \quad (38)$$

Example: Construct Poisson regression

x enters through $\eta = \theta^T x$



We've decided that $Y | X = x \sim \text{Poisson}(\eta)$, what should be the link function f ?

GLM requires that $f(\eta) = \mu = \lambda$, so let's find the canonical link function ψ^{-1} where $\eta = \psi(\mu)$.

$$p(y; \eta) = \exp \log \frac{\lambda^y e^{-\lambda}}{y!} \quad (34)$$

$$= \exp(y \log \lambda - \lambda - \log y!) \quad (35)$$

$$[\text{compare with } h(y) \exp(\eta^T T(y) - A(\eta))] \quad (36)$$

$$(37)$$

$$\eta = \log \lambda \implies f(\theta^T x) = \exp(\theta^T x). \quad (38)$$

Note that $f: \mathbb{R} \rightarrow (0, \infty)$.

Example: multinomial logistic regression

How to extend logistic regression to multiclass classification?

Example: multinomial logistic regression

How to extend logistic regression to multiclass classification?

Response variable: Bernoulli distribution \rightarrow **multinomial distribution**

$$p(y; \pi_1, \dots, \pi_K) = \frac{N!}{y_1! \dots y_K!} \pi_1^{y_1} \dots \pi_K^{y_K} \quad (N = 1) \quad (39)$$

$$\mathbb{E}[Y_k] = p(y_k = 1) = \pi_k \quad (40)$$

Let's find the (canonical) link function following the strategy of Poisson regression.

Multinomial distribution as an exponential family

$$p(y; \pi_1, \dots, \pi_K) = \exp \log \pi_1^{y_1} \dots \pi_K^{y_K} \quad (41)$$

$$= \exp \left(\sum_{k=1}^K y_k \log \pi_k \right) \quad A(\eta) = 0?? \quad (42)$$

(45)

Multinomial distribution as an exponential family

$$p(y; \pi_1, \dots, \pi_K) = \exp \log \pi_1^{y_1} \dots \pi_K^{y_K} \quad (41)$$

$$= \exp \left(\sum_{k=1}^K y_k \log \pi_k \right) \quad A(\eta) = 0?? \quad (42)$$

$$= \exp \left(\sum_{k=1}^{K-1} y_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} y_k \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \quad (43)$$

(45)

Multinomial distribution as an exponential family

$$p(y; \pi_1, \dots, \pi_K) = \exp \log \pi_1^{y_1} \dots \pi_K^{y_K} \quad (41)$$

$$= \exp \left(\sum_{k=1}^K y_k \log \pi_k \right) \quad A(\eta) = 0?? \quad (42)$$

$$= \exp \left(\sum_{k=1}^{K-1} y_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} y_k \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \quad (43)$$

$$= \exp \left(\sum_{k=1}^{K-1} y_k \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} + \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \quad (44)$$

$$(45)$$

Multinomial distribution as an exponential family

$$p(y; \pi_1, \dots, \pi_K) = \exp \log \pi_1^{y_1} \dots \pi_K^{y_K} \quad (41)$$

$$= \exp \left(\sum_{k=1}^K y_k \log \pi_k \right) \quad A(\eta) = 0?? \quad (42)$$

$$= \exp \left(\sum_{k=1}^{K-1} y_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} y_k \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \quad (43)$$

$$= \exp \left(\sum_{k=1}^{K-1} y_k \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} + \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \quad (44)$$

$$= \exp \left(\sum_{k=1}^{K-1} T_k(y) \eta_k - A(\eta) \right) \quad (45)$$

Link function of multinomial logistic regression

$$\eta_k = \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} \quad (46)$$

Link function of multinomial logistic regression

$$\eta_k = \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} = \log \frac{\pi_k}{\pi_K} \quad (46)$$

Link function of multinomial logistic regression

$$\eta_k = \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} = \log \frac{\pi_k}{\pi_K} \quad (46)$$

How do we find the inverse function $f(\eta_k) = \pi_k$?

Link function of multinomial logistic regression

$$\eta_k = \log \frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} = \log \frac{\pi_k}{\pi_K} \quad (46)$$

How do we find the inverse function $f(\eta_k) = \pi_k$?

Exercise: find f by summing π_k .

$$p(y_k = 1 \mid x; \theta) = \pi_k \quad (47)$$

$$= \frac{e^{\eta_k}}{\sum_{k=1}^K e^{\eta_k}} \quad \text{softmax function} \quad (48)$$

$$= \frac{e^{\theta_k^T x}}{\sum_{k=1}^K e^{\theta_k^T x}} \quad (49)$$

Recipe for constructing a GLM:

- 1 Define input and output space (as for any other model).
- 2 Choose the response distribution $Y | X$ that belongs to the exponential family.
- 3 Choose the link function that maps $\theta^T x$ to a proper space.

Next,

- Fit the model by maximum likelihood estimation.
- Closed solutions do not exist in general, so we use gradient ascent.

MLE for GLM

Recall that the likelihood of the exponential family has the form

$$\log p(y \mid x; \theta) = \eta^T y - A(\eta). \quad (50)$$

(53)

MLE for GLM

Recall that the likelihood of the exponential family has the form

$$\log p(y \mid x; \theta) = \eta^T y - A(\eta). \quad (50)$$

Let's compute the gradient of the likelihood.

$$\nabla_{\theta} \ell(\theta) = \sum_{n=1}^N \nabla_{\eta^n} \ell^n \nabla_{\theta} \eta^n \quad \text{chain rule} \quad (51)$$

(53)

MLE for GLM

Recall that the likelihood of the exponential family has the form

$$\log p(y \mid x; \theta) = \eta^T y - A(\eta). \quad (50)$$

Let's compute the gradient of the likelihood.

$$\nabla_{\theta} \ell(\theta) = \sum_{n=1}^N \nabla_{\eta^n} \ell^n \nabla_{\theta} \eta^n \quad \text{chain rule} \quad (51)$$

$$= \sum_{n=1}^N \left(y^{(n)} - \nabla_{\eta^n} A(\eta^n) \right) x^{(n)} \quad \eta^n = \theta^T x^{(n)} \quad (52)$$

$$(53)$$

MLE for GLM

Recall that the likelihood of the exponential family has the form

$$\log p(y \mid x; \theta) = \eta^T y - A(\eta). \quad (50)$$

Let's compute the gradient of the likelihood.

$$\nabla_{\theta} \ell(\theta) = \sum_{n=1}^N \nabla_{\eta^n} \ell^n \nabla_{\theta} \eta^n \quad \text{chain rule} \quad (51)$$

$$= \sum_{n=1}^N \left(y^{(n)} - \nabla_{\eta^n} A(\eta^n) \right) x^{(n)} \quad \eta^n = \theta^T x^{(n)} \quad (52)$$

$$= \sum_{n=1}^N \left(y^{(n)} - \mathbb{E}[Y \mid x^{(n)}; \theta] \right) x^{(n)} \quad A'(\eta) = \mu \quad (53)$$

MLE for GLM

Recall that the likelihood of the exponential family has the form

$$\log p(y | x; \theta) = \eta^T y - A(\eta). \quad (50)$$

Let's compute the gradient of the likelihood.

$$\nabla_{\theta} \ell(\theta) = \sum_{n=1}^N \nabla_{\eta^n} \ell^n \nabla_{\theta} \eta^n \quad \text{chain rule} \quad (51)$$

$$= \sum_{n=1}^N \left(y^{(n)} - \nabla_{\eta^n} A(\eta^n) \right) x^{(n)} \quad \eta^n = \theta^T x^{(n)} \quad (52)$$

$$= \sum_{n=1}^N \left(y^{(n)} - \mathbb{E}[Y | x^{(n)}; \theta] \right) x^{(n)} \quad A'(\eta) = \mu \quad (53)$$

Same form as we've seen in linear regression and logistic regression.

Generative models

We've seen

- Model the conditional distribution $p(y | x; \theta)$ using generalized linear models.
- (Previously) Directly map x to y , e.g., perceptron.

How about

- Model the **joint distribution** $p(x, y; \theta)$.
- Predict the label for x as $\arg \max_{y \in \mathcal{Y}} p(x, y; \theta)$.

Generative modeling through the Bayes rule

Training:

$$p(x, y) \tag{54}$$

(56)

Generative modeling through the Bayes rule

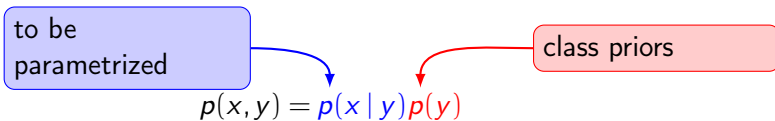
Training:

$$p(x, y) = p(x | y)p(y) \quad (54)$$

(56)

Generative modeling through the Bayes rule

Training:

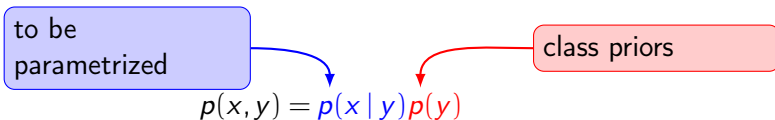


(54)

(56)

Generative modeling through the Bayes rule

Training:



(54)


Testing:

$$p(y|x)$$

(56)

Generative modeling through the Bayes rule

Training:


$$p(x, y) = p(x | y) p(y) \quad (54)$$


Testing:

$$p(y | x) = \frac{p(x | y) p(y)}{p(x)} \quad \text{Bayes rule} \quad (55)$$

(56)

Generative modeling through the Bayes rule

Training:


$$p(x, y) = p(x | y) p(y) \quad (54)$$

Testing:

$$p(y | x) = \frac{p(x | y) p(y)}{p(x)} \quad \text{Bayes rule} \quad (55)$$

$$\arg \max_y p(y | x) = \arg \max_y p(x | y) p(y) \quad (56)$$

Naive Bayes (NB) models

Let's consider binary text classification (e.g., fake vs genuine review) as a motivating example.

Naive Bayes (NB) models

Let's consider binary text classification (e.g., fake vs genuine review) as a motivating example.

Bag-of-words representation of a document

- ["machine", "learning", "is", "fun", "."]
- $x_i \in \{0, 1\}$: whether the i -th word in our vocabulary exists in the input

$$x = [x_1, x_2, \dots, x_d] \quad \text{where } d = \text{vocabulary size} \quad (57)$$

Naive Bayes (NB) models

Let's consider binary text classification (e.g., fake vs genuine review) as a motivating example.

Bag-of-words representation of a document

- ["machine", "learning", "is", "fun", "."]
- $x_i \in \{0, 1\}$: whether the i -th word in our vocabulary exists in the input

$$x = [x_1, x_2, \dots, x_d] \quad \text{where } d = \text{vocabulary size} \quad (57)$$

What's the probability of a document x ?

Naive Bayes (NB) models

Let's consider binary text classification (e.g., fake vs genuine review) as a motivating example.

Bag-of-words representation of a document

- ["machine", "learning", "is", "fun", "."]
- $x_i \in \{0, 1\}$: whether the i -th word in our vocabulary exists in the input

$$x = [x_1, x_2, \dots, x_d] \quad \text{where } d = \text{vocabulary size} \quad (57)$$

What's the probability of a document x ?

$$p(x | y) = p(x_1, \dots, x_d | y) \quad (58)$$

$$= p(x_1 | y) p(x_2 | y, x_1) p(x_3 | y, x_2, x_1) \dots p(x_d | y, x_{d-1}, \dots, x_1) \quad \text{chain rule} \quad (59)$$

$$= \prod_{i=1}^d p(x_i | y, x_{<i}) \quad (60)$$

Naive Bayes assumption

Challenge: $p(x_i | y, x_{<i})$ is hard to model (and estimate), especially for large i .

Naive Bayes assumption

Challenge: $p(x_i | y, x_{<i})$ is hard to model (and estimate), especially for large i .

Solution:

Naive Bayes assumption

Features are **conditionally independent** given the label:

$$p(x | y) = \prod_{i=1}^d p(x_i | y). \quad (61)$$

A strong assumption in generation, but works well in practice.

Parametrize $p(x_i | y)$ and $p(y)$

For binary x_i , assume $p(x_i | y)$ follows Bernoulli distributions.

$$p(x_i = 1 | y = 1) = \theta_{i,1}, \quad p(x_i = 1 | y = 0) = \theta_{i,0}. \quad (62)$$

Parametrize $p(x_i | y)$ and $p(y)$

For binary x_i , assume $p(x_i | y)$ follows Bernoulli distributions.

$$p(x_i = 1 | y = 1) = \theta_{i,1}, \quad p(x_i = 1 | y = 0) = \theta_{i,0}. \quad (62)$$

Similarly,

$$p(y = 1) = \theta_0. \quad (63)$$

Parametrize $p(x_i | y)$ and $p(y)$

For binary x_i , assume $p(x_i | y)$ follows Bernoulli distributions.

$$p(x_i = 1 | y = 1) = \theta_{i,1}, \quad p(x_i = 1 | y = 0) = \theta_{i,0}. \quad (62)$$

Similarly,

$$p(y = 1) = \theta_0. \quad (63)$$

Thus,

$$p(x, y) = p(x | y)p(y) \quad (64)$$

$$= p(y) \prod_{i=1}^d p(x_i | y) \quad \text{NB assumption} \quad (65)$$

$$= p(y) \prod_{i=1}^d \theta_{i,y} \mathbb{I}\{x_i = 1\} + (1 - \theta_{i,y}) \mathbb{I}\{x_i = 0\} \quad (66)$$

Indicator function $\mathbb{I}\{\text{condition}\}$ evaluates to 1 if “condition” is true and 0 otherwise.

MLE for our NB model

We maximize the likelihood of the data $\prod_{n=1}^N p_{\theta}(x^{(n)}, y^{(n)})$ (as opposed to the *conditional* likelihood we've seen before).

(69)

MLE for our NB model

We maximize the likelihood of the data $\prod_{n=1}^N p_{\theta}(x^{(n)}, y^{(n)})$ (as opposed to the *conditional* likelihood we've seen before).

$$\frac{\partial}{\partial \theta_{j,1}} \ell = \frac{\partial}{\partial \theta_j^1} \sum_{n=1}^N \sum_{i=1}^d \log \left(\theta_{i,y^{(n)}} \mathbb{I} \{x_i^{(n)} = 1\} + (1 - \theta_{i,y^{(n)}}) \mathbb{I} \{x_i^{(n)} = 0\} \right) + \log p_{\theta_0}(y^{(n)})$$

(67)

(69)

MLE for our NB model

We maximize the likelihood of the data $\prod_{n=1}^N p_{\theta}(x^{(n)}, y^{(n)})$ (as opposed to the *conditional* likelihood we've seen before).

$$\frac{\partial}{\partial \theta_{j,1}} \ell = \frac{\partial}{\partial \theta_j^1} \sum_{n=1}^N \sum_{i=1}^d \log \left(\theta_{i,y^{(n)}} \mathbb{I} \{x_i^{(n)} = 1\} + (1 - \theta_{i,y^{(n)}}) \mathbb{I} \{x_i^{(n)} = 0\} \right) + \log p_{\theta_0}(y^{(n)}) \quad (67)$$

$$= \frac{\partial}{\partial \theta_{j,1}} \sum_{n=1}^N \log \left(\theta_{j,y^{(n)}} \mathbb{I} \{x_j^{(n)} = 1\} + (1 - \theta_{j,y^{(n)}}) \mathbb{I} \{x_j^{(n)} = 0\} \right) \quad \text{ignore } i \neq j \quad (68)$$

(69)

MLE for our NB model

We maximize the likelihood of the data $\prod_{n=1}^N p_{\theta}(x^{(n)}, y^{(n)})$ (as opposed to the *conditional* likelihood we've seen before).

$$\frac{\partial}{\partial \theta_{j,1}} \ell = \frac{\partial}{\partial \theta_j^1} \sum_{n=1}^N \sum_{i=1}^d \log \left(\theta_{i,y^{(n)}} \mathbb{I} \{x_i^{(n)} = 1\} + (1 - \theta_{i,y^{(n)}}) \mathbb{I} \{x_i^{(n)} = 0\} \right) + \log p_{\theta_0}(y^{(n)}) \quad (67)$$

$$= \frac{\partial}{\partial \theta_{j,1}} \sum_{n=1}^N \log \left(\theta_{j,y^{(n)}} \mathbb{I} \{x_j^{(n)} = 1\} + (1 - \theta_{j,y^{(n)}}) \mathbb{I} \{x_j^{(n)} = 0\} \right) \quad \text{ignore } i \neq j \quad (68)$$

$$= \sum_{n=1}^N \mathbb{I} \{y^{(n)} = 1 \wedge x_j^{(n)} = 1\} \frac{1}{\theta_{j,1}} + \mathbb{I} \{y^{(n)} = 1 \wedge x_j^{(n)} = 0\} \frac{1}{1 - \theta_{j,1}} \quad \text{ignore } y^{(n)} = 0 \quad (69)$$

MLE solution for our NB model

Set $\frac{\partial}{\partial \theta_{j,1}} \ell$ to zero:

$$\theta_{j,1} = \frac{\sum_{n=1}^N \mathbb{I} \left\{ y^{(n)} = 1 \wedge x_j^{(n)} = 1 \right\}}{\sum_{n=1}^N \mathbb{I} \left\{ y^{(n)} = 1 \right\}} \quad (70)$$

MLE solution for our NB model

Set $\frac{\partial}{\partial \theta_{j,1}} \ell$ to zero:

$$\theta_{j,1} = \frac{\sum_{n=1}^N \mathbb{I}\{y^{(n)} = 1 \wedge x_j^{(n)} = 1\}}{\sum_{n=1}^N \mathbb{I}\{y^{(n)} = 1\}} \quad (70)$$

In practice, count words:

$$\frac{\text{number of fake reviews containing "absolutely"}}{\text{number of fake reviews}}$$

Exercise: show that

$$\theta_{j,0} = \frac{\sum_{n=1}^N \mathbb{I}\{y^{(n)} = 0 \wedge x_j^{(n)} = 1\}}{\sum_{n=1}^N \mathbb{I}\{y^{(n)} = 0\}} \quad (71)$$

$$\theta_0 = \frac{\sum_{n=1}^N \mathbb{I}\{y^{(n)} = 1\}}{N} \quad (72)$$

NB assumption: **conditionally independent** features given the label

Recipe for learning a NB model:

- 1 Choose $p(x_i | y)$, e.g., Bernoulli distribution for binary x_i .
- 2 Choose $p(y)$, often a categorical distribution.
- 3 Estimate parameters by MLE (same as the strategy for conditional models) .

Next, NB with continuous features.

NB with continuous inputs

Let's consider a multiclass classification task with continuous inputs.

$$p(x_i | y) \sim \mathcal{N}(\mu_{i,y}, \sigma_{i,y}^2) \quad (73)$$

$$p(y = k) = \theta_k \quad (74)$$

NB with continuous inputs

Let's consider a multiclass classification task with continuous inputs.

$$p(x_i | y) \sim \mathcal{N}(\mu_{i,y}, \sigma_{i,y}^2) \quad (73)$$

$$p(y = k) = \theta_k \quad (74)$$

Likelihood of the data:

$$p(\mathcal{D}) = \prod_{n=1}^N p(y^{(n)}) \prod_{i=1}^d p(x_i^{(n)} | y^{(n)}) \quad (75)$$

$$= \prod_{n=1}^N \theta_{y^{(n)}} \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} \exp\left(-\frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}}\right)^2\right) \quad (76)$$

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^N \log \theta_{y^{(n)}} + \sum_{n=1}^N \sum_{i=1}^d \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \quad (77)$$

(79)

(80)

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^N \log \theta_{y^{(n)}} + \sum_{n=1}^N \sum_{i=1}^d \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \quad (77)$$

$$\frac{\partial}{\partial \mu_{j,k}} \ell = \frac{\partial}{\partial \mu_{j,k}} \sum_{n: y^{(n)}=k} -\frac{1}{2\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right)^2 \quad \text{ignore irrelevant terms} \quad (78)$$

$$(79)$$

$$(80)$$

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^N \log \theta_{y^{(n)}} + \sum_{n=1}^N \sum_{i=1}^d \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \quad (77)$$

$$\frac{\partial}{\partial \mu_{j,k}} \ell = \frac{\partial}{\partial \mu_{j,k}} \sum_{n:y^{(n)}=k} -\frac{1}{2\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right)^2 \quad \text{ignore irrelevant terms} \quad (78)$$

$$= \sum_{n:y^{(n)}=k} \frac{1}{\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right) \quad (79)$$

(80)

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^N \log \theta_{y^{(n)}} + \sum_{n=1}^N \sum_{i=1}^d \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \quad (77)$$

$$\frac{\partial}{\partial \mu_{j,k}} \ell = \frac{\partial}{\partial \mu_{j,k}} \sum_{n:y^{(n)}=k} -\frac{1}{2\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right)^2 \quad \text{ignore irrelevant terms} \quad (78)$$

$$= \sum_{n:y^{(n)}=k} \frac{1}{\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right) \quad (79)$$

Set $\frac{\partial}{\partial \mu_{j,k}} \ell$ to zero:

$$\mu_{j,k} = \frac{\sum_{n:y^{(n)}=k} x_j^{(n)}}{\sum_{n:y^{(n)}=k} 1} \quad (80)$$

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^N \log \theta_{y^{(n)}} + \sum_{n=1}^N \sum_{i=1}^d \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \quad (77)$$

$$\frac{\partial}{\partial \mu_{j,k}} \ell = \frac{\partial}{\partial \mu_{j,k}} \sum_{n:y^{(n)}=k} -\frac{1}{2\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right)^2 \quad \text{ignore irrelevant terms} \quad (78)$$

$$= \sum_{n:y^{(n)}=k} \frac{1}{\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right) \quad (79)$$

Set $\frac{\partial}{\partial \mu_{j,k}} \ell$ to zero:

$$\mu_{j,k} = \frac{\sum_{n:y^{(n)}=k} x_j^{(n)}}{\sum_{n:y^{(n)}=k} 1} = \text{sample mean of } x_j \text{ in class } k \quad (80)$$

Exercise: show that

$$\sigma_{j,k} = \frac{\sum_{n:y^{(n)}=k} \left(x_j^{(n)} - \mu_{j,k}\right)^2}{\sum_{n:y^{(n)}=k} 1} = \text{sample variance of } x_j \text{ in class } k \quad (81)$$

$$\theta_k = \frac{\sum_{n:y^{(n)}=k} 1}{N} \quad (\text{class prior}) \quad (82)$$

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

(87)

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

$$\log \frac{p(y = 1 | x)}{p(y = 0 | x)} = \log \frac{p(x | y = 1)p(y = 1)}{p(x | y = 0)p(y = 0)} \quad (83)$$

(87)

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

$$\log \frac{p(y = 1 | x)}{p(y = 0 | x)} = \log \frac{p(x | y = 1)p(y = 1)}{p(x | y = 0)p(y = 0)} \quad (83)$$

$$= \log \frac{\theta_0}{1 - \theta_0} + \sum_{i=1}^d \left(\log \sqrt{\frac{\sigma_{i,0}^2}{\sigma_{i,1}^2}} + \left(\frac{(x_i - \mu_{i,0})^2}{2\sigma_{i,0}^2} - \frac{(x_i - \mu_{i,1})^2}{2\sigma_{i,1}^2} \right) \right)$$

(87)

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

$$\log \frac{p(y = 1 | x)}{p(y = 0 | x)} = \log \frac{p(x | y = 1)p(y = 1)}{p(x | y = 0)p(y = 0)} \quad (83)$$

$$= \log \frac{\theta_0}{1 - \theta_0} + \sum_{i=1}^d \left(\log \sqrt{\frac{\sigma_{i,0}^2}{\sigma_{i,1}^2}} + \left(\frac{(x_i - \mu_{i,0})^2}{2\sigma_{i,0}^2} - \frac{(x_i - \mu_{i,1})^2}{2\sigma_{i,1}^2} \right) \right) \quad \text{quadratic} \quad (84)$$

(87)

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} \quad (83)$$

$$= \log \frac{\theta_0}{1-\theta_0} + \sum_{i=1}^d \left(\log \sqrt{\frac{\sigma_{i,0}^2}{\sigma_{i,1}^2}} + \left(\frac{(x_i - \mu_{i,0})^2}{2\sigma_{i,0}^2} - \frac{(x_i - \mu_{i,1})^2}{2\sigma_{i,1}^2} \right) \right) \quad \text{quadratic} \quad (84)$$

$$\text{assume that } \sigma_{i,0} = \sigma_{i,1} = \sigma_i, \quad (\theta_0 = 0.5) \quad (85)$$

$$= \sum_{i=1}^d \frac{1}{2\sigma_i^2} \left((x_i - \mu_{i,0})^2 - (x_i - \mu_{i,1})^2 \right) \quad (86)$$

$$= \sum_{i=1}^d \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2} \quad \text{linear} \quad (87)$$

Decision boundary of the Gaussian NB model

Assuming the variance of each feature is the same for both classes, we have

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \sum_{i=1}^d \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2} \quad (88)$$

$$= \theta^T x \quad \text{where else have we seen it?} \quad (89)$$

$$(90)$$

Decision boundary of the Gaussian NB model

Assuming the variance of each feature is the same for both classes, we have

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \sum_{i=1}^d \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2} \quad (88)$$

$$= \theta^T x$$

where else have we seen it? (89)

(90)

$$\theta_i = \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} \quad \text{for } i \in [1, d] \quad (91)$$

$$\theta_0 = \sum_{i=1}^d \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2} \quad \text{bias term} \quad (92)$$

Naive Bayes vs logistic regression

	logistic regression	Gaussian naive Bayes
model type	conditional/discriminative	generative
parametrization	$p(y x)$	$p(x y), p(y)$
assumptions on Y	Bernoulli	Bernoulli
assumptions on X	—	Gaussian
decision boundary	$\theta_{\text{LR}}^T x$	$\theta_{\text{GNB}}^T x$

Naive Bayes vs logistic regression

	logistic regression	Gaussian naive Bayes
model type	conditional/discriminative	generative
parametrization	$p(y x)$	$p(x y), p(y)$
assumptions on Y	Bernoulli	Bernoulli
assumptions on X	—	Gaussian
decision boundary	$\theta_{\text{LR}}^T x$	$\theta_{\text{GNB}}^T x$

Given the same training data, is $\theta_{\text{LR}} = \theta_{\text{GNB}}$?

Naive Bayes vs logistic regression

	logistic regression	GNB
optimization error		
estimation error		
approximation error		

Naive Bayes vs logistic regression

	logistic regression	GNB
optimization error	0 (convex)	0 (closed-form)
estimation error		
approximation error		

Naive Bayes vs logistic regression

	logistic regression	GNB	
optimization error	0 (convex)	0 (closed-form)	
estimation error	0	0	infinite data
approximation error			

Naive Bayes vs logistic regression

	logistic regression	GNB	
optimization error	0 (convex)	0 (closed-form)	
estimation error	0	0	infinite data
approximation error	0	0	GNB assumption holds

Naive Bayes vs logistic regression

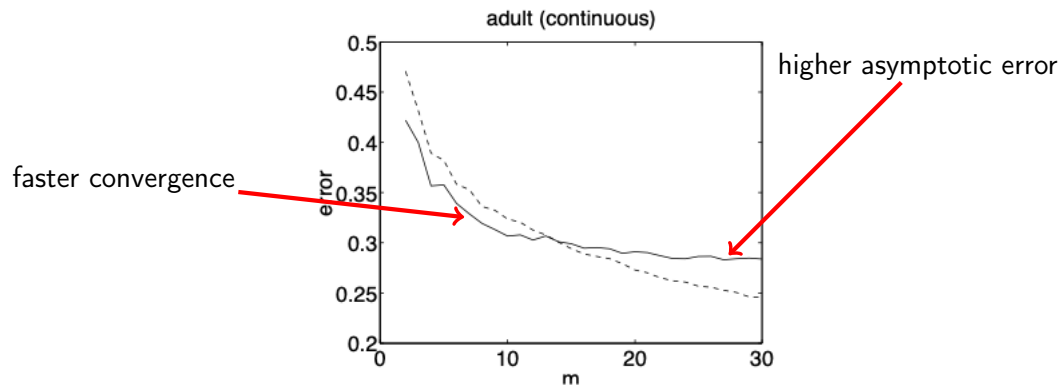
	logistic regression	GNB	
optimization error	0 (convex)	0 (closed-form)	
estimation error	0	0	infinite data
approximation error	0	0	GNB assumption holds

Logistic regression and Gaussian naive Bayes converge to the same classifier asymptotically, assuming the GNB assumption holds.

What if the GNB assumption is not true?

Generative vs discriminative classifiers

Ng, A. and Jordan, M. (2002). [On discriminative versus generative classifiers: A comparison of logistic regression and naive Bayes](#). In Advances in Neural Information Processing Systems 14.



Solid line: naive Bayes; dashed line: logistic regression.