Linear Methods for Classification

TrungTin Nguyen

STATIFY team, Inria centre at the University Grenoble Alpes, France

Statistical Analysis and Document Mining

Complementary Course, Master of Applied Mathematics in Grenoble

Outline

Outline

Previous episode: Nearest-neighbour methods

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \subsetneq \mathbb{N}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Previous episode: Nearest-neighbour methods

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \subsetneq \mathbb{N}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?

Previous episode: Nearest-neighbour methods

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \subsetneq \mathbb{N}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?
- K-Nearest-neighbours (kNN) approximate the probability that Y belongs to a particular category instead Y,

$$\mathbb{P}(Y = c|\mathbf{X} = \mathbf{x}) \approx \frac{1}{K} \sum_{\mathbf{x}_n \in \mathcal{N}_K(\mathbf{x})} \mathbb{1}_{y=c}(y_n).$$

Previous episode: Nearest-neighbour methods

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \subsetneq \mathbb{N}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?
- K-Nearest-neighbours (kNN) approximate the probability that Y belongs to a particular category instead Y,

$$\mathbb{P}(Y = c|\mathbf{X} = \mathbf{x}) \approx \frac{1}{K} \sum_{\mathbf{x}_n \in \mathcal{N}_K(\mathbf{x})} \mathbb{1}_{y=c}(y_n).$$

- kNN is a **nonparametric classifier**: it makes **no assumptions about the data**. $\rightarrow$ It can be **very flexible** but may **not be optimal** when we **know something about the data.**

Previous episode: Nearest-neighbour methods

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \subsetneq \mathbb{N}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?
- K-Nearest-neighbours (kNN) approximate the probability that Y belongs to a particular category instead Y,

$$\mathbb{P}(Y = c|\mathbf{X} = \mathbf{x}) \approx \frac{1}{K} \sum_{\mathbf{x}_n \in \mathcal{N}_K(\mathbf{x})} \mathbb{1}_{y=c}(y_n).$$

- kNN is a **nonparametric classifier**: it makes **no assumptions about the data**. $\to$ It can be **very flexible** but may **not be optimal** when we **know something about the data.**
- The hyper-parameter K usually chosen via cross-validation.

It works well in low dimensions, but suffers from the curse of dimensionality. Verified in CC5!

Outline

Previous episode: high-dimensional data classification

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Previous episode: high-dimensional data classification

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- In genomics and other areas of computational biology, the number of features P is much larger than the number of observations N, often written as $P \gg N$, where $\mathbf{x}_n \in \mathbb{R}^P$.

Previous episode: high-dimensional data classification

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- In genomics and other areas of computational biology, the number of features P is much larger than the number of observations N, often written as $P \gg N$, where $\mathbf{x}_n \in \mathbb{R}^P$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?

Previous episode: high-dimensional data classification

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- In genomics and other areas of computational biology, the number of features P is much larger than the number of observations N, often written as $P \gg N$, where $\mathbf{x}_n \in \mathbb{R}^P$.
- How to model relationship between $p_c(\mathbf{X}) = \mathbb{P}(Y = c|\mathbf{X})$, $c \in \mathcal{C}$, and $\mathbf{X}$?
- Which is suitable for high-dimensional data: Discriminative approaches (CM5) or Generative approaches (CM6).
  1. K-nearest neighbors (K-NN) (CM5),
  2. Logistic Regression (CM5),
  3. Linear Discriminant Analysis (CM6),
  4. Naive Bayes classifier (CM6).

Outline

1. Classification Problems and Curse of Dimensionality
   - Previous episode: nearest-neighbour methods
   - Previous episode: high-dimensional data classification
   - Previous episode: multiple impact of high-dimensionality on statistics
   - Multinomial logistic regression
   - Baseline and softmax coding in multinomial linear regression
2. Generalized Linear Models
   - Previous episode: linear regression for bikeshare data set
   - Bikeshare data: Poisson regression
   - Bikeshare data: linear regression and Poisson regression
   - Generalized linear models
3. A Mathematical Comparison of Classification Methods
   - LDA and multinomial LR
   - LDA, QDA, and naive Bayes

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).
2. The accumulation of small fluctuations in many different directions can produce a large global fluctuation.

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).
2. The accumulation of small fluctuations in many different directions can produce a large global fluctuation.
3. An event that is an accumulation of rare events may not be rare.

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).
2. The accumulation of small fluctuations in many different directions can produce a large global fluctuation.
3. An event that is an accumulation of rare events may not be rare.
4. Numerical computations and optimizations in high-dimensional spaces can be overly intensive.

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).
2. The accumulation of small fluctuations in many different directions can produce a large global fluctuation.
3. An event that is an accumulation of rare events may not be rare.
4. Numerical computations and optimizations in high-dimensional spaces can be overly intensive.

For more details, see [Giraud, 2021, Chapter 1].

Outline

1. Classification Problems and Curse of Dimensionality
   - Previous episode: nearest-neighbour methods
   - Previous episode: high-dimensional data classification
   - Previous episode: multiple impact of high-dimensionality on statistics
   - Multinomial logistic regression
   - Baseline and softmax coding in multinomial linear regression
2. Generalized Linear Models
   - Previous episode: linear regression for bikeshare data set
   - Bikeshare data: Poisson regression
   - Bikeshare data: linear regression and Poisson regression
   - Generalized linear models
3. A Mathematical Comparison of Classification Methods
   - LDA and multinomial LR
   - LDA, QDA, and naive Bayes

Previous episode: multiple logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in [2] \equiv \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Previous episode: multiple logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in [2] \equiv \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- For $k \in [2]$, how to model relationship $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$ and $\mathbf{X}$?

Previous episode: multiple logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in [2] \equiv \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- For $k \in [2]$, how to model relationship $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$ and $\mathbf{X}$?
- Multiple logistic regression takes the form $p_2(\mathbf{X}) = 1 - p_1(\mathbf{X})$, and models the probability that Y belongs to a particular category instead Y,

$$\log\left(\frac{p_1(\mathbf{X})}{1 - p_1(\mathbf{X})}\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p, \text{ or } p_1(\mathbf{X}) = \frac{\exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}{1 + \exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}.$$

Previous episode: multiple logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in [2] \equiv \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- For $k \in [2]$, how to model relationship $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$ and $\mathbf{X}$?
- Multiple logistic regression takes the form $p_2(\mathbf{X}) = 1 - p_1(\mathbf{X})$, and models the probability that Y belongs to a particular category instead Y,

$$\log\left(\frac{p_1(\mathbf{X})}{1 - p_1(\mathbf{X})}\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p, \text{ or } p_1(\mathbf{X}) = \frac{\exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}{1 + \exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}.$$

- **Training error:** using non-linear LS or maximum likelihood estimation (MLE), we obtain $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) = \operatorname{argmax}_{c\in\mathcal{C}} p_c(\mathbf{x}_n)$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[y_n \neq \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right] \approx 0. \quad (1)$$

Previous episode: multiple logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in [2] \equiv \mathcal{C}$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- For $k \in [2]$, how to model relationship $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$ and $\mathbf{X}$?
- Multiple logistic regression takes the form $p_2(\mathbf{X}) = 1 - p_1(\mathbf{X})$, and models the probability that Y belongs to a particular category instead Y,

$$\log\left(\frac{p_1(\mathbf{X})}{1 - p_1(\mathbf{X})}\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p, \text{ or } p_1(\mathbf{X}) = \frac{\exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}{1 + \exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}.$$

- **Training error:** using non-linear LS or maximum likelihood estimation (MLE), we obtain $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) = \operatorname{argmax}_{c\in\mathcal{C}} p_c(\mathbf{x}_n)$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[y_n \neq \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right] \approx 0. \quad (1)$$

Test (generalization) error: for any **new sample** $(\mathbf{x}^*, y^*)$, how we guarantee

$$y^* \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}^*), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}) \equiv \mathbb{E}_{\mathbf{X},Y}\left[\mathbb{1}\left(Y \neq \widehat{r}_{\mathcal{D}}(\mathbf{X})\right)\right] \approx 0? \quad (2)$$

Multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$, $k \in [K]$, and $\mathbf{X}$?

Multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$, $k \in [K]$, and $\mathbf{X}$?
- Multinomial logistic regression (LR) takes the form $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X})$, and models the probability that Y belongs to a particular category instead of the value of Y as follows:

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \text{ or equivalent,}$$

$$p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

Here, we first select a single class to serve as the baseline; without loss of generality, we select the Kth class for this role. It holds that

$$p_K(\mathbf{X}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}. \tag{3}$$

Outline

Multinomial LR: baseline and softmax coding

- Log-odds or logit transformations using baseline coding: $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X}) =?$,

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \quad p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

[1] Jiang and Tanner (1999). On the identifiability of mixtures-of-experts. Neural Networks.

[2] Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification.

Multinomial LR: baseline and softmax coding

- Log-odds or logit transformations using baseline coding: $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X}) = ?$,

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \quad p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

- Log-odds using softmax coding takes the form

$$\log\left(\frac{p_k(\mathbf{X})}{p_k'(\mathbf{X})}\right) = (\beta_{k0} - \beta_{k'0}) + \sum_{p=1}^{P} \left(\beta_{kp} - \beta_{k'p}\right) X_p, \text{ or equivalent,}$$

$$p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

[1] Jiang and Tanner (1999). On the identifiability of mixtures-of-experts. Neural Networks.

[2] Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification.

Multinomial LR: baseline and softmax coding

- Log-odds or logit transformations using baseline coding: $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X}) =?$,

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \quad p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

- Log-odds using softmax coding takes the form

$$\log\left(\frac{p_k(\mathbf{X})}{p'_k(\mathbf{X})}\right) = (\beta_{k0} - \beta_{k'0}) + \sum_{p=1}^{P} \left(\beta_{kp} - \beta_{k'p}\right) X_p, \text{ or equivalent,}$$

$$p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

- Baseline coding avoids identifiability problem, for example in mixture of experts models[1] [2] .

---

[1] Jiang and Tanner (1999). On the identifiability of mixtures-of-experts. Neural Networks.

[2] Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification.

Multinomial LR: baseline and softmax coding

- Log-odds or logit transformations using baseline coding: $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X}) =?$,

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \quad p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

- Log-odds using softmax coding takes the form

$$\log\left(\frac{p_k(\mathbf{X})}{p'_k(\mathbf{X})}\right) = (\beta_{k0} - \beta_{k'0}) + \sum_{p=1}^{P} \left(\beta_{kp} - \beta_{k'p}\right) X_p, \text{ or equivalent,}$$

$$p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

Baseline coding avoids identifiability problem, for example in mixture of experts models[1] [2] .
Softmax coding is used extensively in some areas of the machine learning literature, for example, softmax activation function in deep neural network [James et al., 2021, Chapter 10].

---

[1] Jiang and Tanner (1999). On the identifiability of mixtures-of-experts. Neural Networks.

[2] Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification.

Multinomial LR: baseline and softmax coding

Log-odds or logit transformations using baseline coding: $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X}) =?$,

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p, \quad p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

Log-odds using softmax coding takes the form

$$\log\left(\frac{p_k(\mathbf{X})}{p'_k(\mathbf{X})}\right) = (\beta_{k0} - \beta_{k'0}) + \sum_{p=1}^{P} \left(\beta_{kp} - \beta_{k'p}\right) X_p, \text{ or equivalent,}$$

$$p_k(\mathbf{X}) = \frac{\exp(\beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \sum_{p=1}^{P} \beta_{lp} X_p)}.$$

Baseline coding avoids identifiability problem, for example in mixture of experts models[1] [2] .
Softmax coding is used extensively in some areas of the machine learning literature, for example, softmax activation function in deep neural network [James et al., 2021, Chapter 10].
In multinomial LR, the fitted values, log odds between any pair of classes, and other key ~~model outputs will remain the same, regardless of coding!~~

[1] Jiang and Tanner (1999). On the identifiability of mixtures-of-experts. Neural Networks.

[2] Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification.

Outline

1. Classification Problems and Curse of Dimensionality
   - Previous episode: nearest-neighbour methods
   - Previous episode: high-dimensional data classification
   - Previous episode: multiple impact of high-dimensionality on statistics
   - Multinomial logistic regression
   - Baseline and softmax coding in multinomial linear regression
2. Generalized Linear Models
   - Previous episode: linear regression for bikeshare data set
   - Bikeshare data: Poisson regression
   - Bikeshare data: linear regression and Poisson regression
   - Generalized linear models
3. A Mathematical Comparison of Classification Methods
   - LDA and multinomial LR
   - LDA, QDA, and naive Bayes

Outline

Bikeshare data set: motivation example

☛ We consider the Bikeshare data set. The response is bikers, the **number of hourly users of a bike sharing** program in Washington, DC. This response value is **neither qualitative nor quantitative**: instead, it **takes on non-negative integer values, or counts.**

Bikeshare data set: motivation example

We consider the Bikeshare data set. The response is bikers, the **number of hourly users of a bike sharing** program in Washington, DC. This response value is **neither qualitative nor quantitative**: instead, it **takes on non-negative integer values, or counts.**

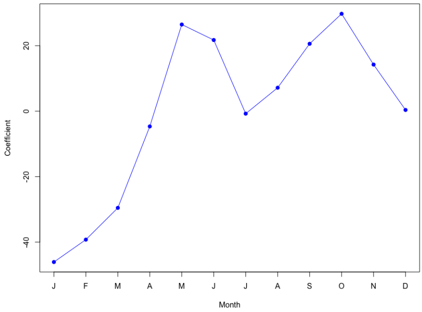
Predicting bikers using the covariates mnth (month of the year), hr (hour of the day, from 0 to 23), workingday (an indicator variable that equals 1 if it is neither a weekend nor a holiday), temp (the normalized temperature, in Celsius), and weathersit (a qualitative variable that takes on one of four possible values: clear; misty or cloudy; light rain or light snow; or heavy rain or heavy snow.)

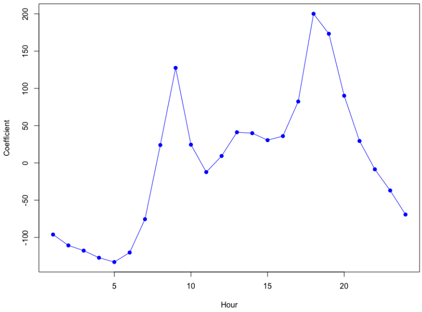
```
> head(Bikeshare)
  season mnth day hr holiday weekday workingday   weathersit temp  atemp  hum windspeed casual registered bikers
1      1  Jan   1  0       0       6          0        clear 0.24 0.2879 0.81    0.0000      3         13     16
2      1  Jan   1  1       0       6          0        clear 0.22 0.2727 0.80    0.0000      8         32     40
3      1  Jan   1  2       0       6          0        clear 0.22 0.2727 0.80    0.0000      5         27     32
4      1  Jan   1  3       0       6          0        clear 0.24 0.2879 0.75    0.0000      3         10     13
5      1  Jan   1  4       0       6          0        clear 0.24 0.2879 0.75    0.0000      0          1      1
6      1  Jan   1  5       0       6          0 cloudy/misty 0.24 0.2576 0.75    0.0896      0          1      1
```

```
> mod.lm2 <- lm(
+     bikers ~ mnth + hr + workingday + temp + weathersit,
+     data = Bikeshare
+   )
> summary(mod.lm2)

Call:
lm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
    data = Bikeshare)

Residuals:
    Min      1Q  Median      3Q     Max
-299.00  -45.70   -6.23   41.08  425.29

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                73.5974     5.1322  14.340  < 2e-16 ***
mnth1                     -46.0871     4.0855 -11.281  < 2e-16 ***
mnth2                     -39.2419     3.5391 -11.088  < 2e-16 ***
mnth3                     -29.5357     3.1552  -9.361  < 2e-16 ***
mnth4                      -4.6622     2.7406  -1.701  0.08895 .
mnth5                      26.4700     2.8508   9.285  < 2e-16 ***
mnth6                      21.7317     3.4651   6.272 3.75e-10 ***
mnth7                      -0.7626     3.9084  -0.195  0.84530
mnth8                       7.1560     3.5347   2.024  0.04295 *
mnth9                      20.5912     3.0456   6.761 1.46e-11 ***
mnth10                     29.7472     2.6995  11.019  < 2e-16 ***
mnth11                     14.2229     2.8604   4.972 6.74e-07 ***
```

```
mnth11                  14.2229     2.8604   4.972 6.74e-07 ***
hr1                    -96.1420     3.9554 -24.307  < 2e-16 ***
hr2                   -110.7213     3.9662 -27.916  < 2e-16 ***
hr3                   -117.7212     4.0165 -29.310  < 2e-16 ***
hr4                   -127.2828     4.0808 -31.191  < 2e-16 ***
hr5                   -133.0495     4.1168 -32.319  < 2e-16 ***
hr6                   -120.2775     4.0370 -29.794  < 2e-16 ***
hr7                    -75.5424     3.9916 -18.925  < 2e-16 ***
hr8                     23.9511     3.9686   6.035 1.65e-09 ***
hr9                    127.5199     3.9500  32.284  < 2e-16 ***
hr10                    24.4399     3.9360   6.209 5.57e-10 ***
hr11                   -12.3407     3.9361  -3.135  0.00172 **
hr12                     9.2814     3.9447   2.353  0.01865 *
hr13                    41.1417     3.9571  10.397  < 2e-16 ***
hr14                    39.8939     3.9750  10.036  < 2e-16 ***
hr15                    30.4940     3.9910   7.641 2.39e-14 ***
hr16                    35.9445     3.9949   8.998  < 2e-16 ***
hr17                    82.3786     3.9883  20.655  < 2e-16 ***
hr18                   200.1249     3.9638  50.488  < 2e-16 ***
hr19                   173.2989     3.9561  43.806  < 2e-16 ***
hr20                    90.1138     3.9400  22.872  < 2e-16 ***
hr21                    29.4071     3.9362   7.471 8.74e-14 ***
hr22                    -8.5883     3.9332  -2.184  0.02902 *
hr23                   -37.0194     3.9344  -9.409  < 2e-16 ***
workingday               1.2696     1.7845   0.711  0.47681
temp                   157.2094    10.2612  15.321  < 2e-16 ***
weathersitcloudy/misty -12.8903     1.9643  -6.562 5.60e-11 ***
weathersitlight rain/snow -66.4944  2.9652 -22.425  < 2e-16 ***
```

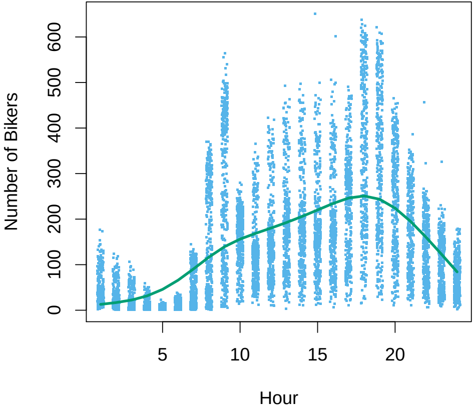
```
hr20                         90.1138     3.9400  22.872  < 2e-16 ***
hr21                         29.4071     3.9362   7.471 8.74e-14 ***
hr22                         -8.5883     3.9332  -2.184  0.02902 *
hr23                        -37.0194     3.9344  -9.409  < 2e-16 ***
workingday                    1.2696     1.7845   0.711  0.47681
temp                        157.2094    10.2612  15.321  < 2e-16 ***
weathersitcloudy/misty      -12.8903     1.9643  -6.562 5.60e-11 ***
weathersitlight rain/snow   -66.4944     2.9652 -22.425  < 2e-16 ***
weathersitheavy rain/snow -109.7446    76.6674  -1.431  0.15234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.5 on 8605 degrees of freedom
Multiple R-squared:  0.6745,	Adjusted R-squared:  0.6731
F-statistic: 457.3 on 39 and 8605 DF,  p-value: < 2.2e-16
```



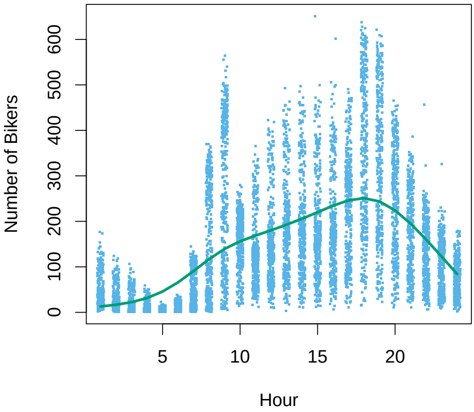


A least squares linear regression model was fit to predict bikers in the Bikeshare data set. Left: The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter. Right: The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight [James et al., 2021, Figure 4.13]

❓ At first glance, fitting a linear regression model to the Bikeshare **seems to provide reasonable and intuitive results.**



For the most part, as the **mean number of bikers increases**, so does **the variance in the number of bikers** [James et al., 2021, Figure 4.14].



For the most part, as the **mean number of bikers increases**, so does **the variance in the number of bikers** [James et al., 2021, Figure 4.14].

❓ At first glance, fitting a linear regression model to the Bikeshare **seems to provide reasonable and intuitive results.**

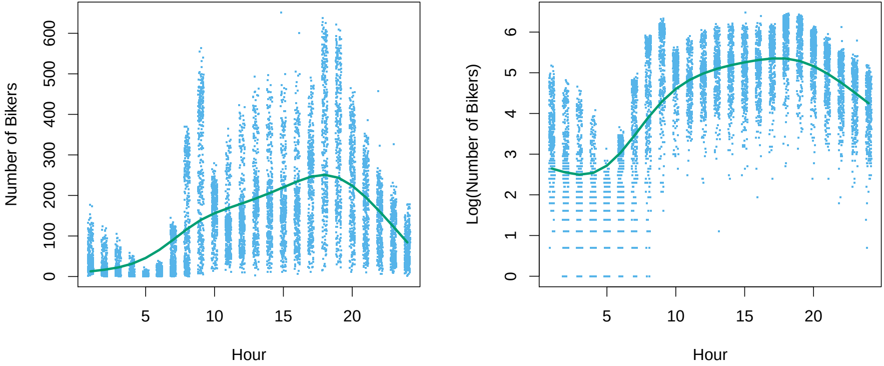
Some issues

1. **9.6% of the fitted values** in the Bikeshare data set are **negative**: that is, the linear regression model predicts a negative number of users during 9.6% of the hours in the data set.

Some issues

1. **9.6% of the fitted values** in the Bikeshare data set are **negative**: that is, the linear regression model predicts a negative number of users during 9.6% of the hours in the data set.
2. Since ϵ is a continuous-valued error term, response Y is necessarily continuous-valued (quantitative) but the response bikers is integer-valued.

3. The mean-variance relationship is a major violation of the assumptions of a linear model, which state that $Y = \beta_0 + \sum_{p=1}^{P} X_p \beta_p + \epsilon$, where ϵ is a mean-zero error term with variance σ^2 that is constant, and not a function of the covariates. For the most part, as the mean number of bikers increases, so does the variance in the number of bikers!



Previous episode: some issues for Bikeshare data set using linear regression

1. Transforming the response Y avoids the possibility of negative predictions, and it overcomes much of the heteroscedasticity in the untransformed data $\log(Y) = \beta_0 + \sum_{p=1}^{P} X_p\beta_p + \epsilon$.

Previous episode: some issues for Bikeshare data set using linear regression

1. Transforming the response Y avoids the possibility of negative predictions, and it overcomes much of the heteroscedasticity in the untransformed data $\log(Y) = \beta_0 + \sum_{p=1}^{P} X_p \beta_p + \epsilon$.
   BUT Making predictions and inference in terms of the log of the response, rather than the response $\rightarrow$challenges in interpretation!

Previous episode: some issues for Bikeshare data set using linear regression

1. Transforming the response Y avoids the possibility of negative predictions, and it overcomes much of the heteroscedasticity in the untransformed data $\log(Y) = \beta_0 + \sum_{p=1}^{P} X_p \beta_p + \epsilon$.
   BUT Making predictions and inference in terms of the log of the response, rather than the response $\rightarrow$challenges in interpretation!

Poisson regression model provides a much more natural and elegant approach for this task! (To be verified soon!)

Previous episode: some issues for Bikeshare data set using linear regression

1. Transforming the response Y avoids the possibility of negative predictions, and it overcomes much of the heteroscedasticity in the untransformed data $\log(Y) = \beta_0 + \sum_{p=1}^{P} X_p\beta_p + \epsilon$.
   BUT Making predictions and inference in terms of the log of the response, rather than the response $\rightarrow$challenges in interpretation!

- Poisson regression model provides a much more natural and elegant approach for this task! (To be verified soon!)
- Transforming the mean of response $\mathbb{E}[Y]$ as in the logistic regression!

Outline

Poisson Regression

- Definition of Poisson distribution:

- Definition of Poisson distribution: Suppose that a random variable Y takes on nonnegative integer values, Poisson, *i.e.,* $Y \in \{0, 1, 2, ...\}$. If Y follows the Poisson distribution, then

$$\mathbb{P}(Y = k) = \frac{e^{\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots, \lambda > 0. \tag{4}$$

Here, $\mathbb{E}[Y] = \mathrm{var}[Y] =$

Poisson Regression

- Definition of Poisson distribution: Suppose that a random variable Y takes on nonnegative integer values, Poisson, *i.e.,* $Y \in \{0, 1, 2, ...\}$. If Y follows the Poisson distribution, then

$$\mathbb{P}(Y = k) = \frac{e^{\lambda}\lambda^{k}}{k!} \text{ for } k = 0, 1, 2, \ldots, \lambda > 0. \tag{4}$$

  Here, $\mathbb{E}[Y] = \text{var}[Y] = \lambda$.

- Definition of Poisson regression model:

Poisson Regression

- Definition of Poisson distribution: Suppose that a random variable Y takes on nonnegative integer values, Poisson, *i.e.,* $Y \in \{0, 1, 2, ...\}$. If Y follows the Poisson distribution, then

$$\mathbb{P}(Y = k) = \frac{e^{\lambda} \lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots, \lambda > 0. \tag{4}$$

  Here, $\mathbb{E}[Y] = \text{var}[Y] = \lambda$.

- Definition of Poisson regression model:

$$\mathbb{E}[Y] = \lambda \equiv \lambda(X_1, \ldots, X_P) = \exp(\beta_0 + \sum_{p=1}^{P} X_p \beta_p). \tag{5}$$

Poisson Regression

- Definition of Poisson distribution: Suppose that a random variable Y takes on nonnegative integer values, Poisson, *i.e.,* $Y \in \{0, 1, 2, ...\}$. If Y follows the Poisson distribution, then

$$\mathbb{P}(Y = k) = \frac{e^{\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots, \lambda > 0. \tag{4}$$

  Here, $\mathbb{E}[Y] = \mathrm{var}[Y] = \lambda$.

- Definition of Poisson regression model:

$$\mathbb{E}[Y] = \lambda \equiv \lambda(X_1, \ldots, X_P) = \exp(\beta_0 + \sum_{p=1}^{P} X_p\beta_p). \tag{5}$$

- MLE approach: we want to maximize MLE

$$l(\beta_0, \beta_1, \ldots, \beta_P) = \prod_{n=1}^{N} \frac{e^{-\lambda(x_n)}\lambda(x_n)^{y_n}}{y_n!}, \quad \lambda(x_n) = \exp(\beta_0 + \sum_{p=1}^{P} x_{np}\beta_p). \tag{6}$$

Outline

1. Classification Problems and Curse of Dimensionality
   - Previous episode: nearest-neighbour methods
   - Previous episode: high-dimensional data classification
   - Previous episode: multiple impact of high-dimensionality on statistics
   - Multinomial logistic regression
   - Baseline and softmax coding in multinomial linear regression
2. Generalized Linear Models
   - Previous episode: linear regression for bikeshare data set
   - Bikeshare data: Poisson regression
   - Bikeshare data: linear regression and Poisson regression
   - Generalized linear models
3. A Mathematical Comparison of Classification Methods
   - LDA and multinomial LR
   - LDA, QDA, and naive Bayes

```
> mod.lm2 <- lm(
+     bikers ~ mnth + hr + workingday + temp + weathersit,
+     data = Bikeshare
+   )
> summary(mod.lm2)

Call:
lm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
    data = Bikeshare)

Residuals:
    Min      1Q  Median      3Q     Max 
-299.00  -45.70   -6.23   41.08  425.29 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 73.5974     5.1322  14.340  < 2e-16 ***
mnth1                      -46.0871     4.0855 -11.281  < 2e-16 ***
mnth2                      -39.2419     3.5391 -11.088  < 2e-16 ***
mnth3                      -29.5357     3.1552  -9.361  < 2e-16 ***
mnth4                       -4.6622     2.7406  -1.701  0.08895 .  
mnth5                       26.4700     2.8508   9.285  < 2e-16 ***
mnth6                       21.7317     3.4651   6.272 3.75e-10 ***
mnth7                       -0.7626     3.9084  -0.195  0.84530    
mnth8                        7.1560     3.5347   2.024  0.04295 *  
mnth9                       20.5912     3.0456   6.761 1.46e-11 ***
mnth10                      29.7472     2.6995  11.019  < 2e-16 ***
mnth11                      14.2229     2.8604   4.972 6.74e-07 ***
```

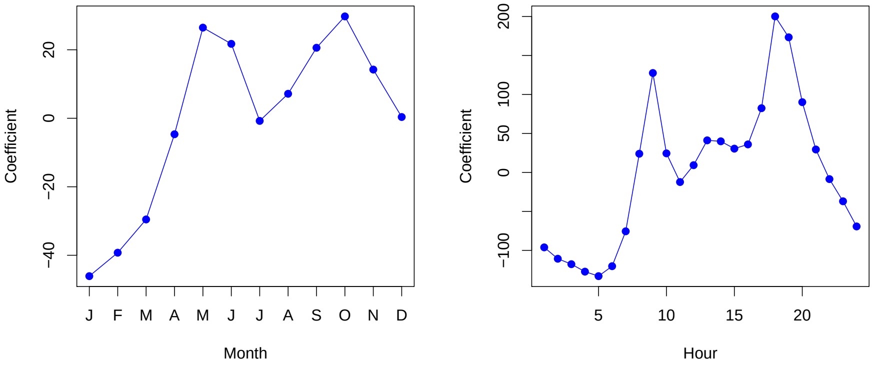
```
> mod.pois <- glm(
+     bikers ~ mnth + hr + workingday + temp + weathersit,
+     data = Bikeshare, family = poisson
+   )
> summary(mod.pois)

Call:
glm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
    family = poisson, data = Bikeshare)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.7574   -3.3441   -0.6549    2.6999   21.9628

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                4.118245   0.006021  683.964  < 2e-16 ***
mnth1                     -0.670170   0.005907 -113.445  < 2e-16 ***
mnth2                     -0.444124   0.004860  -91.379  < 2e-16 ***
mnth3                     -0.293733   0.004144  -70.886  < 2e-16 ***
mnth4                      0.021523   0.003125    6.888 5.66e-12 ***
mnth5                      0.240471   0.002916   82.462  < 2e-16 ***
mnth6                      0.223235   0.003554   62.818  < 2e-16 ***
mnth7                      0.103617   0.004125   25.121  < 2e-16 ***
mnth8                      0.151171   0.003662   41.281  < 2e-16 ***
mnth9                      0.233493   0.003102   75.281  < 2e-16 ***
mnth10                     0.267573   0.002785   96.091  < 2e-16 ***
mnth11                     0.150264   0.003180   47.248  < 2e-16 ***
```

```
mnth11                      14.2229     2.8604   4.972 6.74e-07 ***
hr1                        -96.1420     3.9554 -24.307  < 2e-16 ***
hr2                       -110.7213     3.9662 -27.916  < 2e-16 ***
hr3                       -117.7212     4.0165 -29.310  < 2e-16 ***
hr4                       -127.2828     4.0808 -31.191  < 2e-16 ***
hr5                       -133.0495     4.1168 -32.319  < 2e-16 ***
hr6                       -120.2775     4.0370 -29.794  < 2e-16 ***
hr7                        -75.5424     3.9916 -18.925  < 2e-16 ***
hr8                         23.9511     3.9686   6.035 1.65e-09 ***
hr9                        127.5199     3.9500  32.284  < 2e-16 ***
hr10                        24.4399     3.9360   6.209 5.57e-10 ***
hr11                       -12.3407     3.9361  -3.135  0.00172 **
hr12                         9.2814     3.9447   2.353  0.01865 *
hr13                        41.1417     3.9571  10.397  < 2e-16 ***
hr14                        39.8939     3.9750  10.036  < 2e-16 ***
hr15                        30.4940     3.9910   7.641 2.39e-14 ***
hr16                        35.9445     3.9949   8.998  < 2e-16 ***
hr17                        82.3786     3.9883  20.655  < 2e-16 ***
hr18                       200.1249     3.9638  50.488  < 2e-16 ***
hr19                       173.2989     3.9561  43.806  < 2e-16 ***
hr20                        90.1138     3.9400  22.872  < 2e-16 ***
hr21                        29.4071     3.9362   7.471 8.74e-14 ***
hr22                        -8.5883     3.9332  -2.184  0.02902 *
hr23                       -37.0194     3.9344  -9.409  < 2e-16 ***
workingday                   1.2696     1.7845   0.711  0.47681
temp                       157.2094    10.2612  15.321  < 2e-16 ***
weathersitcloudy/misty     -12.8903     1.9643  -6.562 5.60e-11 ***
weathersitlight rain/snow  -66.4944     2.9652 -22.425  < 2e-16 ***
```

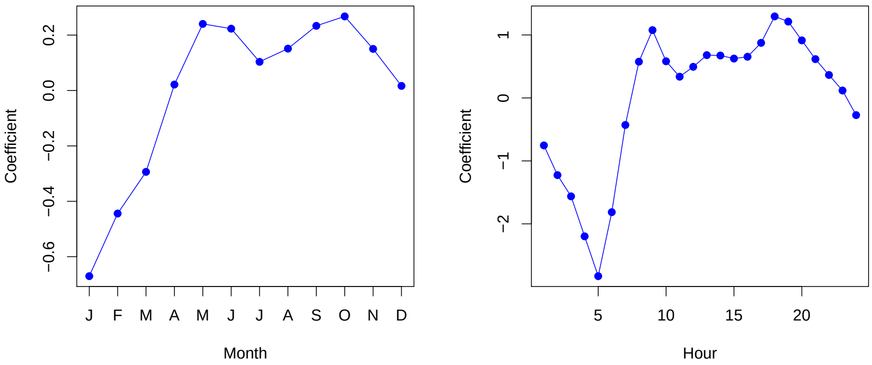
```
hr11                       0.336852   0.004720   71.372  < 2e-16 ***
hr12                       0.494121   0.004392  112.494  < 2e-16 ***
hr13                       0.679642   0.004069  167.040  < 2e-16 ***
hr14                       0.673565   0.004089  164.722  < 2e-16 ***
hr15                       0.624910   0.004178  149.570  < 2e-16 ***
hr16                       0.653763   0.004132  158.205  < 2e-16 ***
hr17                       0.874301   0.003784  231.040  < 2e-16 ***
hr18                       1.294635   0.003254  397.848  < 2e-16 ***
hr19                       1.212281   0.003321  365.084  < 2e-16 ***
hr20                       0.914022   0.003700  247.065  < 2e-16 ***
hr21                       0.616201   0.004191  147.045  < 2e-16 ***
hr22                       0.364181   0.004659   78.173  < 2e-16 ***
hr23                       0.117493   0.005225   22.488  < 2e-16 ***
workingday                 0.014665   0.001955    7.502 6.27e-14 ***
temp                       0.785292   0.011475   68.434  < 2e-16 ***
weathersitcloudy/misty    -0.075231   0.002179  -34.528  < 2e-16 ***
weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1052921  on 8644  degrees of freedom
Residual deviance:  228041  on 8605  degrees of freedom
AIC: 281159

Number of Fisher Scoring iterations: 5
```

```
hr20                          90.1138      3.9400  22.872  < 2e-16 ***
hr21                          29.4071      3.9362   7.471 8.74e-14 ***
hr22                          -8.5883      3.9332  -2.184  0.02902 *
hr23                         -37.0194      3.9344  -9.409  < 2e-16 ***
workingday                     1.2696      1.7845   0.711  0.47681
temp                         157.2094     10.2612  15.321  < 2e-16 ***
weathersitcloudy/misty       -12.8903      1.9643  -6.562 5.60e-11 ***
weathersitlight rain/snow    -66.4944      2.9652 -22.425  < 2e-16 ***
weathersitheavy rain/snow -109.7446     76.6674  -1.431  0.15234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.5 on 8605 degrees of freedom
Multiple R-squared:  0.6745,	Adjusted R-squared:  0.6731
F-statistic: 457.3 on 39 and 8605 DF,  p-value: < 2.2e-16
```

```
hr20                       0.914022   0.003700  247.065  < 2e-16 ***
hr21                       0.616201   0.004191  147.045  < 2e-16 ***
hr22                       0.364181   0.004659   78.173  < 2e-16 ***
hr23                       0.117493   0.005225   22.488  < 2e-16 ***
workingday                 0.014665   0.001955    7.502 6.27e-14 ***
temp                       0.785292   0.011475   68.434  < 2e-16 ***
weathersitcloudy/misty    -0.075231   0.002179  -34.528  < 2e-16 ***
weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1052921  on 8644  degrees of freedom
Residual deviance:  228041  on 8605  degrees of freedom
AIC: 281159

Number of Fisher Scoring iterations: 5
```



A least squares linear regression model was fit to predict bikers in the Bikeshare data set. Left: The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter. Right: The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight [James et al., 2021, Figure 4.13]



A Poisson regression model was fit to predict bikers in the Bikeshare data set. Left: The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter. Right: The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight [James et al., 2021, Figure 4.15]

Important distinction: Poisson and linear regression models

1. Coefficient associated with workingday is **statistically significant** under the Poisson regression model, but not under the linear regression model. **More realistic modeling!**

Important distinction: Poisson and linear regression models

1. Coefficient associated with workingday is **statistically significant** under the Poisson regression model, but not under the linear regression model. **More realistic modeling!**
2. **Mean-variance relationship:** in Poisson regression, we implicitly assume that **mean bike usage in a given hour equals the variance of bike usage during that hour** while a constant variance in linear regression model.
3. **Nonnegative fitted values:** there are no negative predictions using the Poisson regression model.

Outline

1. Classification Problems and Curse of Dimensionality
   - Previous episode: nearest-neighbour methods
   - Previous episode: high-dimensional data classification
   - Previous episode: multiple impact of high-dimensionality on statistics
   - Multinomial logistic regression
   - Baseline and softmax coding in multinomial linear regression
2. Generalized Linear Models
   - Previous episode: linear regression for bikeshare data set
   - Bikeshare data: Poisson regression
   - Bikeshare data: linear regression and Poisson regression
   - Generalized linear models
3. A Mathematical Comparison of Classification Methods
   - LDA and multinomial LR
   - LDA, QDA, and naive Bayes

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, \ldots, X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1 + e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1 + e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, \ldots, X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1 + e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, \dots, X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1 + e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

HOW.

Generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

HOW. $\eta(\mu) = \mu$, $\eta(\mu) = \log(\mu/(1-\mu))$, $\eta(\mu) = \log(\mu)$.

Outline

Outline

Log odds of posterior probabilities with normal assumptions

We now make an analytical (or mathematical) comparison between Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), naive Bayes and multinomial LR, see [James et al., 2021] for more details.

Log odds of posterior probabilities with normal assumptions

We now make an analytical (or mathematical) comparison between Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), naive Bayes and multinomial LR, see [James et al., 2021] for more details.

- We consider these approaches in a setting with K classes, so that we assign an observation to the class that maximizes $\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})$.

Log odds of posterior probabilities with normal assumptions

We now make an analytical (or mathematical) comparison between Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), naive Bayes and multinomial LR, see [James et al., 2021] for more details.

- We consider these approaches in a setting with K classes, so that we assign an observation to the class that maximizes $\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})$.
- Equivalently, via considering K as the baseline class, Bayes' Theorem and $\mathbf{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma)$, we aim to maximize

$$\log\left(\frac{\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = K|\mathbf{X} = \mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p \quad (7)$$

- **What are the value of a_k and b_{kp}?**

Log odds of posteriors in LDA and multinomial LR

In LDA, we maximize the following log odds of the posterior:

$$
\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K) \\
&= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv a_k} + \mathbf{x}^\top \underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv \boldsymbol{b}_k} \\
&= a_k + \sum_{p=1}^{P} b_{kp} x_p, \ \text{linear in } \mathbf{x}. \qquad (8)
\end{aligned}
$$

Log odds of posteriors in LDA and multinomial LR

In LDA, we maximize the following log odds of the posterior:

$$\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right)\\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right)\\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K)\\
&= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv a_k} + \mathbf{x}^\top \underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv \boldsymbol{b}_k}\\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}. \qquad (8)
\end{aligned}$$

Recall that for multinomial LR:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \beta_{k0} + \sum_{p=1}^{P}\beta_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Log odds of posteriors in LDA and multinomial LR

In LDA, we maximize the following log odds of the posterior:

$$
\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_K) \\
&= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv a_k} + \mathbf{x}^\top \underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)}_{\equiv \boldsymbol{b}_k} \\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p, \ \text{linear in } \mathbf{x}. \qquad (8)
\end{aligned}
$$

Recall that for multinomial LR:

$$
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp}x_p, \ \text{linear in } \mathbf{x}.
$$

Both LDA and multinomial LR assume that the log odds of the posterior probabilities is linear in $\mathbf{x}$.

Outline

Log odds of posterior probabilities in QDA and LDA

In QDA, $\mathbf{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ we maximize the following log odds of the posterior:

$$\begin{aligned}
\log\left(\frac{\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = K|\mathbf{X} = \mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y = k)\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = k)}{\mathbb{P}(Y = K)\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x} - \boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x} - \boldsymbol{\mu}_K) \\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_px_q, \ \text{quadratic in } \mathbf{x}, \qquad (9)
\end{aligned}$$

Log odds of posterior probabilities in QDA and LDA

In QDA, $\mathbf{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ we maximize the following log odds of the posterior:

$$
\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K) \\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_p x_q, \text{ quadratic in } \mathbf{x}, \qquad (9)
\end{aligned}
$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$.

Log odds of posterior probabilities in QDA and LDA

In QDA, $\mathbf{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ we maximize the following log odds of the posterior:

$$
\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K) \\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_p x_q, \text{ quadratic in } \mathbf{x}, \qquad (9)
\end{aligned}
$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.
$$

Log odds of posterior probabilities in QDA and LDA

In QDA, $\mathbf{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ we maximize the following log odds of the posterior:

$$
\begin{aligned}
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) &= \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) \\
&= \log\left(\frac{\pi_k \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}{\pi_K \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K))}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_K)^\top \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K) \\
&= a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_p x_q, \ \text{quadratic in } \mathbf{x}, \qquad (9)
\end{aligned}
$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$
\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \ \text{linear in } \mathbf{x}.
$$

LDA is a special case of QDA. This is not surprising, since LDA is simply a restricted version of QDA with $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\sigma}_K = \boldsymbol{\Sigma}$.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (10)$$

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model}, \tag{10}$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (10)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp} x_p, \text{ linear in } \mathbf{x}.$$

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (10)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kp}(x_p) = b_{kp}x_p$.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{10}$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kp}(x_p) = b_{kp}x_p$. This means that this means that LDA is a special case of naive Bayes!

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (10)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kp}(x_p) = b_{kp}x_p$. This means that this means that LDA is a special case of naive Bayes!

This is a little bit surprising!

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{10}$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp} x_p, \text{ linear in } \mathbf{x}.$$

Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kp}(x_p) = b_{kp}x_p$. This means that this means that LDA is a special case of naive Bayes!

This is a little bit surprising!

Each method makes very different assumptions:

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{\mathbb{P}(Y=k)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=k)}{\mathbb{P}(Y=K)\mathbb{P}(\mathbf{X}=\mathbf{x}|Y=K)}\right) = \log\left(\frac{\pi_k \prod_{p=1}^{P} f_{kp}(x_p)}{\pi_K \prod_{p=1}^{P} f_{Kp}(x_p))}\right)$$

$$= \underbrace{\log\left(\frac{\pi_k}{\pi_K}\right)}_{\equiv a_k} + \sum_{p=1}^{P} \underbrace{\log\left(\frac{f_{kp}(x_p)}{f_{Kp}(x_p))}\right)}_{\equiv g_{kp}(x_p)} = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (10)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kp}(x_p) = b_{kp}x_p$. This means that this means that LDA is a special case of naive Bayes!

This is a little bit surprising!

Each method makes very different assumptions: LDA assumes that the features are normally distributed with a common within-class covariance matrix, and naive Bayes instead assumes independence of the features.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{11}$$

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = K|\mathbf{X} = \mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{11}$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (11)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (11)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

If we model $f_{kp}(x_p) = \mathcal{N}(\mu_{kp}, \sigma_p^2)$, then we end up with $g_{kp}(x_p) = b_{kp}x_p$, where $b_{kp} \equiv (\mu_{kp} - \mu_{Kp})/\sigma_p^2$. In this case, naive Bayes, is actually a special case of LDA with $\boldsymbol{\Sigma}$ restricted to be a diagonal matrix with pth diagonal element equal to σ_p^2.

Log odds of posterior probabilities in naive Bayes and LDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (11)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for LDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p, \text{ linear in } \mathbf{x}.$$

If we model $f_{kp}(x_p) = \mathcal{N}(\mu_{kp}, \sigma_p^2)$, then we end up with $g_{kp}(x_p) = b_{kp}x_p$, where $b_{kp} \equiv (\mu_{kp} - \mu_{Kp})/\sigma_p^2$. In this case, naive Bayes, is actually a special case of LDA with $\boldsymbol{\Sigma}$ restricted to be a diagonal matrix with pth diagonal element equal to σ_p^2.

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = K|\mathbf{X} = \mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{12}$$

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$.

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \tag{12}$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_p x_q, \text{ quadratic in } \mathbf{x}.$$

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_p x_q, \text{ quadratic in } \mathbf{x}.$$

Naive Bayes can produce a more flexible fit, since any choice can be made for $g_{kp}(x_p)$. It is restricted to a purely additive fit, a function of x_p is added to a function of x_q, for $p \neq q$; however, these terms are never multiplied!

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_px_q, \text{ quadratic in } \mathbf{x}.$$

Naive Bayes can produce a more flexible fit, since any choice can be made for $g_{kp}(x_p)$. It is restricted to a purely additive fit, a function of x_p is added to a function of x_q, for $p \neq q$; however, these terms are never multiplied!
QDA includes multiplicative terms of the form $c_{kpq}x_px_q$.

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_px_q, \text{ quadratic in } \mathbf{x}.$$

Naive Bayes can produce a more flexible fit, since any choice can be made for $g_{kp}(x_p)$. It is restricted to a purely additive fit, a function of x_p is added to a function of x_q, for $p \neq q$; however, these terms are never multiplied!

QDA includes multiplicative terms of the form $c_{kpq}x_px_q$.

Therefore, QDA has the potential to be more accurate in settings where interactions among the predictors are important in discriminating between classes.

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_px_q, \text{ quadratic in } \mathbf{x}.$$

Naive Bayes can produce a more flexible fit, since any choice can be made for $g_{kp}(x_p)$. It is restricted to a purely additive fit, a function of x_p is added to a function of x_q, for $p \neq q$; however, these terms are never multiplied!
QDA includes multiplicative terms of the form $c_{kpq}x_px_q$.
Therefore, QDA has the potential to be more accurate in settings where interactions among the predictors are important in discriminating between classes.
Neither QDA nor naive Bayes is a special case of the other!

Log odds of posterior probabilities in naive Bayes and QDA

In naive Bayes setting, $f_k(\mathbf{x}) = \prod_{p=1}^{P} f_{kp}(x_p)$, we maximize the following log odds:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} g_{kp}(x_p), \text{ generalized additive model,} \quad (12)$$

where a_k, b_{kp}, c_{kpq} are functions of $\pi_k, \pi_K, \boldsymbol{\mu}_k, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_K$. Recall that for QDA:

$$\log\left(\frac{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}\right) = a_k + \sum_{p=1}^{P} b_{kp}x_p + \sum_{p=1}^{P}\sum_{q=1}^{Q} c_{kpq}x_px_q, \text{ quadratic in } \mathbf{x}.$$

Naive Bayes can produce a more flexible fit, since any choice can be made for $g_{kp}(x_p)$. It is restricted to a purely additive fit, a function of x_p is added to a function of x_q, for $p \neq q$; however, these terms are never multiplied!

QDA includes multiplicative terms of the form $c_{kpq}x_px_q$.

Therefore, QDA has the potential to be more accurate in settings where interactions among the predictors are important in discriminating between classes.

Neither QDA nor naive Bayes is a special case of the other!

References

Giraud, C. (2021).
Introduction to High-Dimensional Statistics, volume 2 of *Monographs on Statistics & Applied Probability*.
Taylor & Francis.

(Cited on pages 15, 16, 17, 18, and 19.)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).
An Introduction to Statistical Learning: with Applications in R, volume 2 of *Springer Texts in Statistics*.
Springer.

(Cited on pages 30, 31, 32, 33, 34, 42, 43, 44, 65, 66, 81, 82, and 83.)