Statistica II (750AA)

Lezione 5

Dario Trevisan – *https://web.dm.unipi.it/trevisan*

20/10/2025

Formula di Bayes e Classificazione

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:
  - $P(\theta|X = x)$: Probabilità a posteriori del parametro θ, date le caratteristiche $X = x$.

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:
  - $P(\theta|X = x)$: Probabilità a posteriori del parametro θ, date le caratteristiche $X = x$.
  - $P(X = x|\theta) = \mathcal{L}(\theta; X = x)$: verosimiglianza del parametro θ relativa alle osservazioni X.

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:
  - $P(\theta|X = x)$: Probabilità a posteriori del parametro θ, date le caratteristiche $X = x$.
  - $P(X = x|\theta) = \mathcal{L}(\theta; X = x)$: verosimiglianza del parametro θ relativa alle osservazioni X.
  - $P(\theta)$: Probabilità a priori del parametro.

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:
  - $P(\theta|X = x)$: Probabilità a posteriori del parametro θ, date le caratteristiche $X = x$.
  - $P(X = x|\theta) = \mathcal{L}(\theta; X = x)$: verosimiglianza del parametro θ relativa alle osservazioni X.
  - $P(\theta)$: Probabilità a priori del parametro.
  - $P(X)$: Probabilità marginale delle caratteristiche.

Introduzione alla Formula di Bayes

- **Formula di Bayes**: date variabili aleatorie $X \in F$ *features*, $\theta \in \Theta$ (parametri, *labels*)

$$P(\theta|X = x) = \frac{P(\theta) \cdot P(X = x|\theta)}{P(X = x)}$$

- **Interpretazione**:
  - $P(\theta|X = x)$: Probabilità a posteriori del parametro θ, date le caratteristiche $X = x$.
  - $P(X = x|\theta) = \mathcal{L}(\theta; X = x)$: verosimiglianza del parametro θ relativa alle osservazioni X.
  - $P(\theta)$: Probabilità a priori del parametro.
  - $P(X)$: Probabilità marginale delle caratteristiche.
- **Osservazione**: $P(X)$ è comune a tutti i parametri θ:

$$P(\theta|X = x) \propto P(\theta)\mathcal{L}(\theta; X = x)$$

Stimatori (puntuali)

Data una densità (es., a posteriori) per θ possiamo determinarne la *moda*:

- **Stima Massimo a Posteriori (MAP)**:

$$\theta_{MAP} := \arg\max_{\theta} P(\theta|X = x) = \arg\max_{\theta} P(\theta)\mathcal{L}(\theta; X = x)$$

Stimatori (puntuali)

Data una densità (es., a posteriori) per θ possiamo determinarne la *moda*:

- **Stima Massimo a Posteriori (MAP)**:

$$\theta_{MAP} := \arg\max_{\theta} P(\theta|X = x) = \arg\max_{\theta} P(\theta)\mathcal{L}(\theta; X = x)$$

- **Stima di Massima Verosimiglianza (MLE)**:

$$\theta_{MLE} := \arg\max_{\theta} P(X = x|\theta) = \arg\max_{\theta} \mathcal{L}(\theta; X = x)$$

densità unif a priori per θ

Stimatori (puntuali)

Data una densità (es., a posteriori) per θ possiamo determinarne la *moda*:

- **Stima Massimo a Posteriori (MAP)**:

$$\theta_{MAP} := \arg\max_{\theta} P(\theta|X = x) = \arg\max_{\theta} P(\theta)\mathcal{L}(\theta; X = x)$$

- **Stima di Massima Verosimiglianza (MLE)**:

$$\theta_{MLE} := \arg\max_{\theta} P(X = x|\theta) = \arg\max_{\theta} \mathcal{L}(\theta; X = x)$$

- Nella *densità* è rappresentata più informazione rispetto alla stima puntuale.

Esempio 1: Stima del parametro di una Bernoulli

$X \in \{0, 1\}$

- **Problema**: Stimare la frazione $\theta \in [0, 1]$ di palline rosse in un'urna effettuando *estrazioni con rimpiazzo*.

Esempio 1: Stima del parametro di una Bernoulli

- **Problema**: Stimare la frazione $\theta \in [0, 1]$ di palline rosse in un'urna effettuando *estrazioni con rimpiazzo*.
- **Modello**:

Esempio 1: Stima del parametro di una Bernoulli

- **Problema**: Stimare la frazione $\theta \in [0, 1]$ di palline rosse in un'urna effettuando *estrazioni con rimpiazzo*.
- **Modello**:
  - $X =$ numero di rosse in n estrazioni

$$\mathcal{L}(\theta; X = x) = P(X = x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Esempio 1: Stima del parametro di una Bernoulli

- **Problema**: Stimare la frazione $\theta \in [0, 1]$ di palline rosse in un'urna effettuando *estrazioni con rimpiazzo*.
- **Modello**:
  - $X =$ numero di rosse in n estrazioni

$$\mathcal{L}(\theta; X = x) = P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

  - Densità a priori uniforme $P(\theta) = 1$ per $\theta \in [0, 1]$

$$\Rightarrow \theta_{MAP} = \theta_{MLE} = \frac{x}{n}.$$

$$\theta_{MLE} = \theta_{MAP} \in \arg\max_{\theta} \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$$0 = \frac{d}{d\theta}\left(\theta^x(1-\theta)^{n-x}\right) = x\theta^{x-1}(1-\theta)^{n-x} - \theta^x(n-x)(1-\theta)^{n-x-1} \dots$$

Angry Bayes

- Giocate voi! https://dario-trevisan.shinyapps.io/AngryBayes/

Angry Bayes

- Giocate voi! https://dario-trevisan.shinyapps.io/AngryBayes/
- Codice (R shiny) disponibile nel team e nella pagina del corso

Densità beta

- Definiamo in generale Beta($\alpha > 0, \beta > 0$) una densità continua

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Densità beta

- Definiamo in generale Beta($\alpha > 0, \beta > 0$) una densità continua

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- *Intepretazione*: $\alpha - 1$ è il numero di successi e $\beta - 1$ è il numero di insuccessi.

$$p(\theta \mid X = x) \propto \theta^{x}(1-\theta)^{n-x}$$

Densità beta

- Definiamo in generale Beta($\alpha > 0, \beta > 0$) una densità continua

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- *Intepretazione*: $\alpha - 1$ è il numero di successi e $\beta - 1$ è il numero di insuccessi.
- La *moda* è $\theta^* = \frac{\alpha-1}{\beta+\alpha-2}$.

Densità beta

- Definiamo in generale Beta($\alpha > 0, \beta > 0$) una densità continua

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- *Intepretazione*: $\alpha - 1$ è il numero di successi e $\beta - 1$ è il numero di insuccessi.
- La *moda* è $\theta^* = \frac{\alpha-1}{\beta+\alpha-2}$.
- La *media* è $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$ $= \int_0^1 \theta\, p(\theta)\, d\theta$

Densità beta

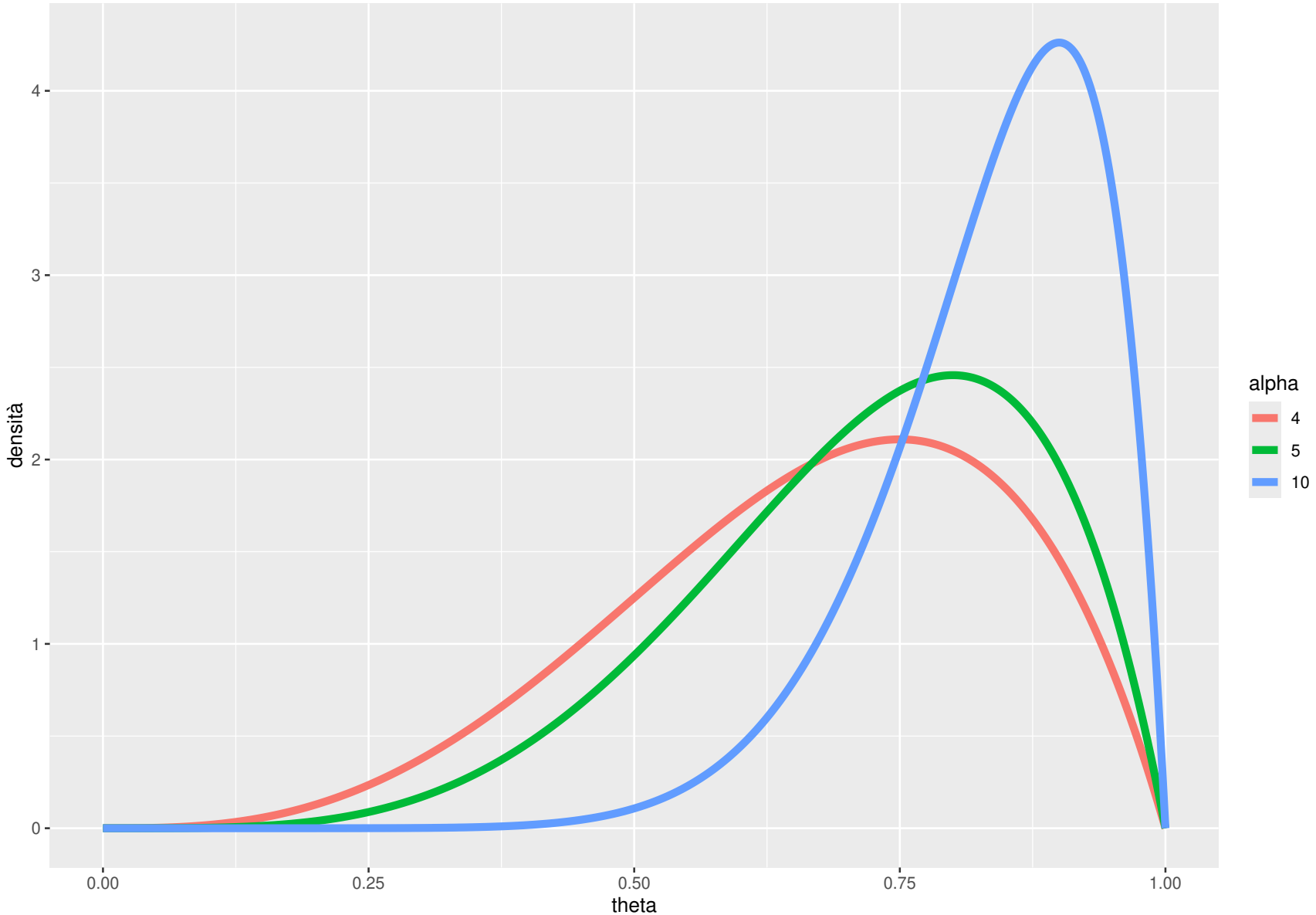
- Definiamo in generale Beta($\alpha > 0, \beta > 0$) una densità continua

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- *Intepretazione*: $\alpha - 1$ è il numero di successi e $\beta - 1$ è il numero di insuccessi.
- La *moda* è $\theta^* = \frac{\alpha-1}{\beta+\alpha-2}$.
- La *media* è $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$
- La *varianza* è $\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \to 0$ se $\alpha + \beta \to \infty$

Esempio.

Densità beta con parametri $\alpha \in \{4, 5, 10\}$ e $\beta = 2$.



Previsione

- **Problema**: avendo osservato in n estrazioni consecutive r palline rosse e $n - r$ blu, qual è la probabilità che nell' estrazione successiva si osservino ~~k rosse~~ 1 rossa?

Previsione

- **Problema**: avendo osservato in n estrazioni consecutive r palline rosse e $n - r$ blu, qual è la probabilità che nell' estrazione successiva si osservino k rosse?
- **Due approcci**:

Previsione

- **Problema**: avendo osservato in n estrazioni consecutive r palline rosse e $n - r$ blu, qual è la probabilità che nell' estrazione successiva si osservino k rosse?

- **Due approcci**:

  1. **Approssimato**: usiamo la stima puntuale $\theta_{MLE} = r/n$ e quindi

$$P(X = 1|X_n = r) \approx r/n.$$

Previsione

- **Problema**: avendo osservato in n estrazioni consecutive r palline rosse e $n - r$ blu, qual è la probabilità che nell' estrazione successiva si osservino k rosse?

- **Due approcci**:

  1. **Approssimato**: usiamo la stima puntuale $\theta_{MLE} = r/n$ e quindi

$$P(X = 1|X_n = r) \approx r/n.$$

  2. *Esatto*: usiamo la legge della probabilità totale (disintegrazione):

$$P(X = 1|X_n = r) = \int_0^1 P(X = 1|\theta, X_n = r)P(\theta|X_n = r)d\theta$$

Da cui la **regola di successione di Laplace**:

$$P(X = 1|X_n = r) = \frac{r+1}{n+2}$$

Smoothing di Laplace

Esempio 2: Media e varianza di una Gaussiana

- **Problema**: Stimare la media e la deviazione standard di una distribuzione gaussiana reale da osservazioni $(X_i)_{i=1}^n$ indipendenti.

Esempio 2: Media e varianza di una Gaussiana

- **Problema**: Stimare la media e la deviazione standard di una distribuzione gaussiana reale da osservazioni $(X_i)_{i=1}^n$ indipendenti.
- **Modello**: $\theta = (m, \sigma)$

Esempio 2: Media e varianza di una Gaussiana

- **Problema**: Stimare la media e la deviazione standard di una distribuzione gaussiana reale da osservazioni $(X_i)_{i=1}^n$ indipendenti.
- **Modello**: $\theta = (m, \sigma)$
  - $\{X = x\} = \{X_i = x_i\}_{i=1}^n$

$$\mathcal{L}(\theta; X = x) = P(X = x|\theta) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right) \frac{1}{\sqrt{2\pi\sigma^2}}.$$

Esempio 2: Media e varianza di una Gaussiana

- **Problema**: Stimare la media e la deviazione standard di una distribuzione gaussiana reale da osservazioni $(X_i)_{i=1}^n$ indipendenti.
- **Modello**: $\theta = (m, \sigma)$
  - $\{X = x\} = \{X_i = x_i\}_{i=1}^n$

$$\mathcal{L}(\theta; X = x) = P(X = x|\theta) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right)\frac{1}{\sqrt{2\pi\sigma^2}}.$$

  - Densità a priori *uniforme* $P(m, \sigma) = \frac{1}{\sigma}$ per $m \in \mathbb{R}, \sigma > 0$

Esempio 2: Media e varianza di una Gaussiana

- **Problema**: Stimare la media e la deviazione standard di una distribuzione gaussiana reale da osservazioni $(X_i)_{i=1}^n$ indipendenti.
- **Modello**: $\theta = (m, \sigma)$
  - $\{X = x\} = \{X_i = x_i\}_{i=1}^n$

$$\mathcal{L}(\theta; X = x) = P(X = x|\theta) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right)\frac{1}{\sqrt{2\pi\sigma^2}}.$$

  - Densità a priori *uniforme* $P(m, \sigma) = \frac{1}{\sigma}$ per $m \in \mathbb{R}, \sigma > 0$
- **Densità a posteriori**:

$$P(m, \sigma|X = x) \propto \exp\left(-\frac{(m - \bar{x}_n)^2}{2(\sigma/\sqrt{n})^2} - \frac{\mathrm{Var}\,(x)_n^2}{2(\sigma/\sqrt{n})^2}\right)\sigma^{-(n+1)}$$

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

media campionaria

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}\,(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Stima MAP : $m_{MAP} = \bar{x}_n$, $\sigma^2_{MAP} = \frac{n}{n+1}\mathrm{Var}\,(x)_n$.

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}\,(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Stima MAP : $m_{MAP} = \bar{x}_n$, $\sigma^2_{MAP} = \frac{n}{n+1}\,\mathrm{Var}\,(x)_n$.
- Previsione:

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}\,(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Stima MAP : $m_{MAP} = \bar{x}_n$, $\sigma^2_{MAP} = \frac{n}{n+1}\,\mathrm{Var}\,(x)_n$.
- Previsione:
  1. **Approssimata** (usando stima *MLE*):

$$P(X_{n+1} = x_{n+1} | X_1 = x_1 \ldots, X_n = x_n) = \mathcal{N}(\bar{x}_n, \mathrm{Var}\,(x)_n)$$

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Stima MAP : $m_{MAP} = \bar{x}_n$, $\sigma^2_{MAP} = \frac{n}{n+1}\mathrm{Var}(x)_n$.
- Previsione:
  1. **Approssimata** (usando stima *MLE*):

$$P(X_{n+1} = x_{n+1} | X_1 = x_1 \ldots, X_n = x_n) = \mathcal{N}(\bar{x}_n, \mathrm{Var}(x)_n)$$

  2. Bayesiana assumendo σ^2 noto:

$$P(X_{n+1} = x_{n+1} | X_1 = x_1 \ldots, X_n = x_n) = \mathcal{N}\left(\bar{x}_n, \frac{n+1}{n}\sigma^2\right)$$

Stime e previsione

- Stima MLE: $m_{MLE} = \bar{x}_n$, $\sigma^2_{MLE} = \mathrm{Var}\,(x)_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Stima MAP : $m_{MAP} = \bar{x}_n$, $\sigma^2_{MAP} = \frac{n}{n+1}\,\mathrm{Var}\,(x)_n$.
- Previsione:
  1. **Approssimata** (usando stima *MLE*):

$$P(X_{n+1} = x_{n+1} | X_1 = x_1 \ldots, X_n = x_n) = \mathcal{N}(\bar{x}_n, \mathrm{Var}\,(x)_n)$$

  2. Bayesiana assumendo σ^2 noto:

$$P(X_{n+1} = x_{n+1} | X_1 = x_1 \ldots, X_n = x_n) = \mathcal{N}\left(\bar{x}_n, \frac{n+1}{n}\sigma^2\right)$$

  3. **Bayesiana esatta**: Student-t con $n-1$ gradi di libertà:

$$P(X_{n+1} = x_{n+1}) \propto \left(1 + \frac{1}{n+1}\frac{(x_{n+1} - \bar{x}_n)^2}{\mathrm{Var}\,(x)_n}\right)^{-n/2}$$

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.
- **Modelli Discriminativi**: $P(C|X)$

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.
- **Modelli Discriminativi**: $P(C|X)$
  - Esempio: Regressione logistica., K-NN

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.
- **Modelli Discriminativi**: $P(C|X)$
  - Esempio: Regressione logistica.
- La formula di Bayes li collega:

$$P(C|X) \propto P(C)\overbrace{P(X|C)}^{\mathcal{L}}$$

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.
- **Modelli Discriminativi**: $P(C|X)$
  - Esempio: Regressione logistica.
- La formula di Bayes li collega:

$$P(C|X) \propto P(C)P(X|C)$$

- Classificatore: $h(x) = \arg\max_\ell P(C = \ell | X = x)$

Modelli di classificazione

- **Modelli Generativi**: $P(X|C)$
  - Esempio: Naive Bayes, LDA e QDA.
- **Modelli Discriminativi**: $P(C|X)$
  - Esempio: Regressione logistica.
- La formula di Bayes li collega:

$$P(C|X) \propto P(C)P(X|C)$$

- Classificatore: $h(x) = \arg\max_\ell P(C = \ell|X = x)$
- Si *modellizzano* e poi si *stimano* le probabilità $P(C|X)$, $P(X|C)$ e $P(C)$ a partire dalle osservazioni del *training set*.

Modello Naive Bayes

- **Problema:** modellizzare $P(X|C)$ quando $X = (X_j)_{j=1}^{d} \in \mathbb{R}^d$ con d elevato

Modello Naive Bayes

- **Problema:** modellizzare $P(X|C)$ quando $X = (X_j)_{j=1}^{d} \in \mathbb{R}^d$ con d elevato
- soluzione *idiota* (ma efficace): ipotesi di indipendenza!

$$P(X|C) = \prod_{j=1}^{d} P(X_j|C)$$

Modello Naive Bayes

- **Problema:** modellizzare $P(X|C)$ quando $X = (X_j)_{j=1}^d \in \mathbb{R}^d$ con d elevato
- soluzione *idiota* (ma efficace): ipotesi di indipendenza!

$$P(X|C) = \prod_{j=1}^{d} P(X_j|C)$$

- **Due esempi**:

Modello Naive Bayes

- **Problema:** modellizzare $P(X|C)$ quando $X = (X_j)_{j=1}^d \in \mathbb{R}^d$ con d elevato
- soluzione *idiota* (ma efficace): ipotesi di indipendenza!

$$P(X|C) = \prod_{j=1}^{d} P(X_j|C)$$

- **Due esempi**:
  - **Bernoulli Naive Bayes**: per dati binari.

Modello Naive Bayes

- **Problema:** modellizzare $P(X|C)$ quando $X = (X_j)_{j=1}^d \in \mathbb{R}^d$ con d elevato
- soluzione *idiota* (ma efficace): ipotesi di indipendenza!

$$P(X|C) = \prod_{j=1}^{d} P(X_j|C)$$

- **Due esempi**:
  - **Bernoulli Naive Bayes**: per dati binari.
  - **Gaussian Naive Bayes**: per dati continui.

Bernoulli Naive Bayes

Supponiamo che ciascuna caratteristica sia *binaria* (es, presente/assente, vero/falso, ecc.).

- Possiamo *modellizzare*:
  $P(X_j = x_j | C = c) = p(j|c)^{x_j}(1 - p(j|c))^{1-x_j}$

Bernoulli Naive Bayes

Supponiamo che ciascuna caratteristica sia *binaria* (es, presente/assente, vero/falso, ecc.).

- Possiamo *modellizzare*:
  $P(X_j = x_j | C = c) = p(j|c)^{x_j}(1 - p(j|c))^{1-x_j}$
- Parametri del modello: $(p(j|c))_{j,c} \in [0,1]^{d \times \sharp C}$,
  $p(c) = P(C = c) \in [0,1]^{\sharp C}$.

Bernoulli Naive Bayes: addestramento e previsione

- *Addestramento* (training) tramite stima dei parametri (MAP): $p(c) = n(c)/n$ e

$$p(j|c) = \frac{n(j|c)}{n(c)} = \frac{\text{num. di oss. con label } c \text{ e feature } j \text{ presente}}{\text{num. di oss. con label } c}$$

Bernoulli Naive Bayes: addestramento e previsione

- *Addestramento* (training) tramite stima dei parametri (MAP): $p(c) = n(c)/n$ e

$$p(j|c) = \frac{n(j|c)}{n(c)} = \frac{\text{num. di oss. con label } c \text{ e feature } j \text{ presente}}{\text{num. di oss. con label } c}$$

- Si può aggiungere un parametro $\alpha > 0$ (prior non uniforme) per evitare $p(j|c) \in \{0, 1\}$:

$$p(j|c) = \frac{n(j|c) + \alpha}{n(c) + 2\alpha}.$$

Bernoulli Naive Bayes: addestramento e previsione

- *Addestramento* (training) tramite stima dei parametri (MAP): $p(c) = n(c)/n$ e

$$p(j|c) = \frac{n(j|c)}{n(c)} = \frac{\text{num. di oss. con label } c \text{ e feature } j \text{ presente}}{\text{num. di oss. con label } c}$$

- Si può aggiungere un parametro $\alpha > 0$ (prior non uniforme) per evitare $p(j|c) \in \{0, 1\}$:

$$p(j|c) = \frac{n(j|c) + \alpha}{n(c) + 2\alpha}.$$

- *Classificatore*: dato un test point con caratteristiche x si calcolano le probabilità delle labels:

$$P(C = c|X = x) \propto p(c) \prod_{j=1}^{d} p(j|c)^{x_j} (1 - p(j|c))^{1-x_j}$$

e si pone $h(x) = c$ la *moda*.

Caso binario $C = \{-1, 1\}$.

- Possiamo usare le *odds*

$$\frac{P(C = +1|X = x)}{P(C = -1|X = x)} = \frac{p(+1)\prod_{j=1}^{d} p(j|+1)^{x_j}(1 - p(j|+1))^{1-x_j}}{p(-1)\prod_{j=1}^{d} p(j|-1)^{x_j}(1 - p(j|-1))^{1-x_j}}$$

oppure passare alle log-odds (*logits*).

Caso binario $C = \{-1, 1\}$.

- Possiamo usare le *odds*

$$\frac{P(C=+1|X=x)}{P(C=-1|X=x)} = \frac{p(+1)\prod_{j=1}^{d} p(j|+1)^{x_j}(1-p(j|+1))^{1-x_j}}{p(-1)\prod_{j=1}^{d} p(j|-1)^{x_j}(1-p(j|-1))^{1-x_j}}$$

oppure passare alle log-odds (*logits*).

- Funzione di punteggio (**score**) lineare nelle features:

$$s(x) := \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

$$\beta_j := \log(p(j|+1)/p(j|-1)) - \log((1-p(j|+1))/(1-p(j|-1)))$$

$$\beta_0 := \log(p(+1)/p(-1)) + \sum_{j=1}^{d} \log((1-p(j|+1))/(1-p(j|-1)))$$

Caso binario $C = \{-1, 1\}$.

- Possiamo usare le *odds*

$$\frac{P(C = +1|X = x)}{P(C = -1|X = x)} = \frac{p(+1)\prod_{j=1}^{d} p(j|+1)^{x_j}(1 - p(j|+1))^{1-x_j}}{p(-1)\prod_{j=1}^{d} p(j|-1)^{x_j}(1 - p(j|-1))^{1-x_j}}$$

  oppure passare alle log-odds (*logits*).

- Funzione di punteggio (**score**) lineare nelle features:

$$s(x) := \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

$$\beta_j := \log(p(j|+1)/p(j|-1)) - \log((1-p(j|+1))/(1-p(j|-1)))$$

$$\beta_0 := \log(p(+1)/p(-1)) + \sum_{j=1}^{d} \log((1-p(j|+1))/(1-p(j|-1)))$$

- Il classificatore è $h(x) = \text{sign}(s(x))$.

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:
  - $s(x)$ indica log-odds $\to \text{Odds}(C = 1|X = x) > e^t$ per $h_t(x) = 1$.

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:
  - $s(x)$ indica log-odds $\to$ $\text{Odds}(C = 1 | X = x) > e^t$ per $h_t(x) = 1$.
  - β_0 diventa $\beta_0 - t$ $\to$ $\text{Odds}(+1) := e^{-t} p(+1)/(p(-1))$

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:
  - $s(x)$ indica log-odds $\to$ Odds$(C = 1|X = x) > e^t$ per $h_t(x) = 1$.
  - β_0 diventa $\beta_0 - t \to$ Odds$(+1) := e^{-1}p(+1)/(p(-1))$
  - Per $t \to \infty$, $h_t(x) = -1 \to FPR = 0$, $TPR = 0$

Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:
  - $s(x)$ indica log-odds $\to$ Odds$(C = 1|X = x) > e^t$ per $h_t(x) = 1$.
  - β_0 diventa $\beta_0 - t \to$ Odds$(+1) := e^{-1}p(+1)/(p(-1))$
  - Per $t \to \infty$, $h_t(x) = -1 \to FPR = 0$, $TPR = 0$
  - Per $t \to -\infty$, $h_t(x) = +1$

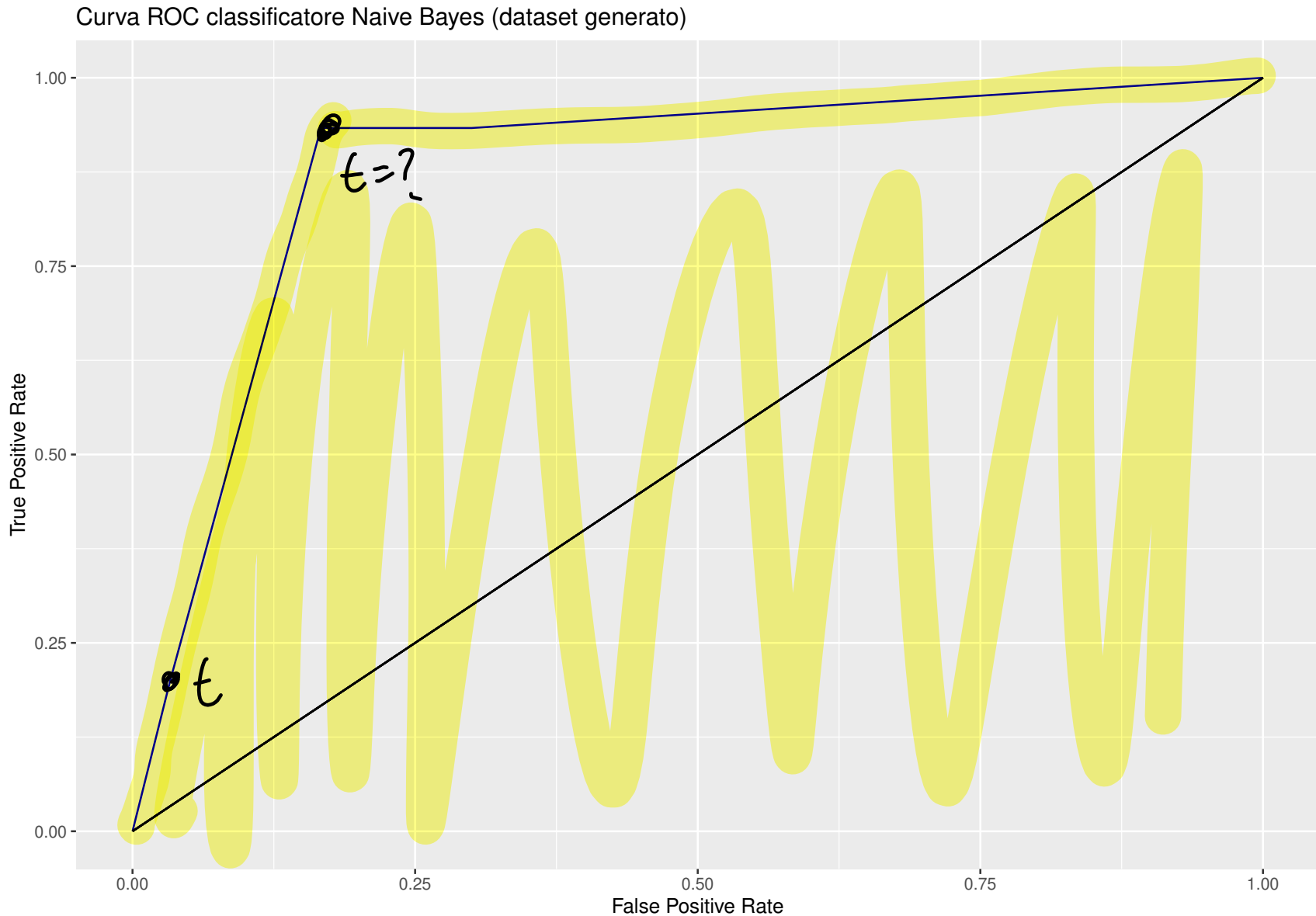
Curva ROC

- Dato uno *score* $s : F \to \mathbb{R}$ e una *soglia* (threshold) $t \in \mathbb{R}$ definiamo un classificatore ($C = \{-1, 1\}$):

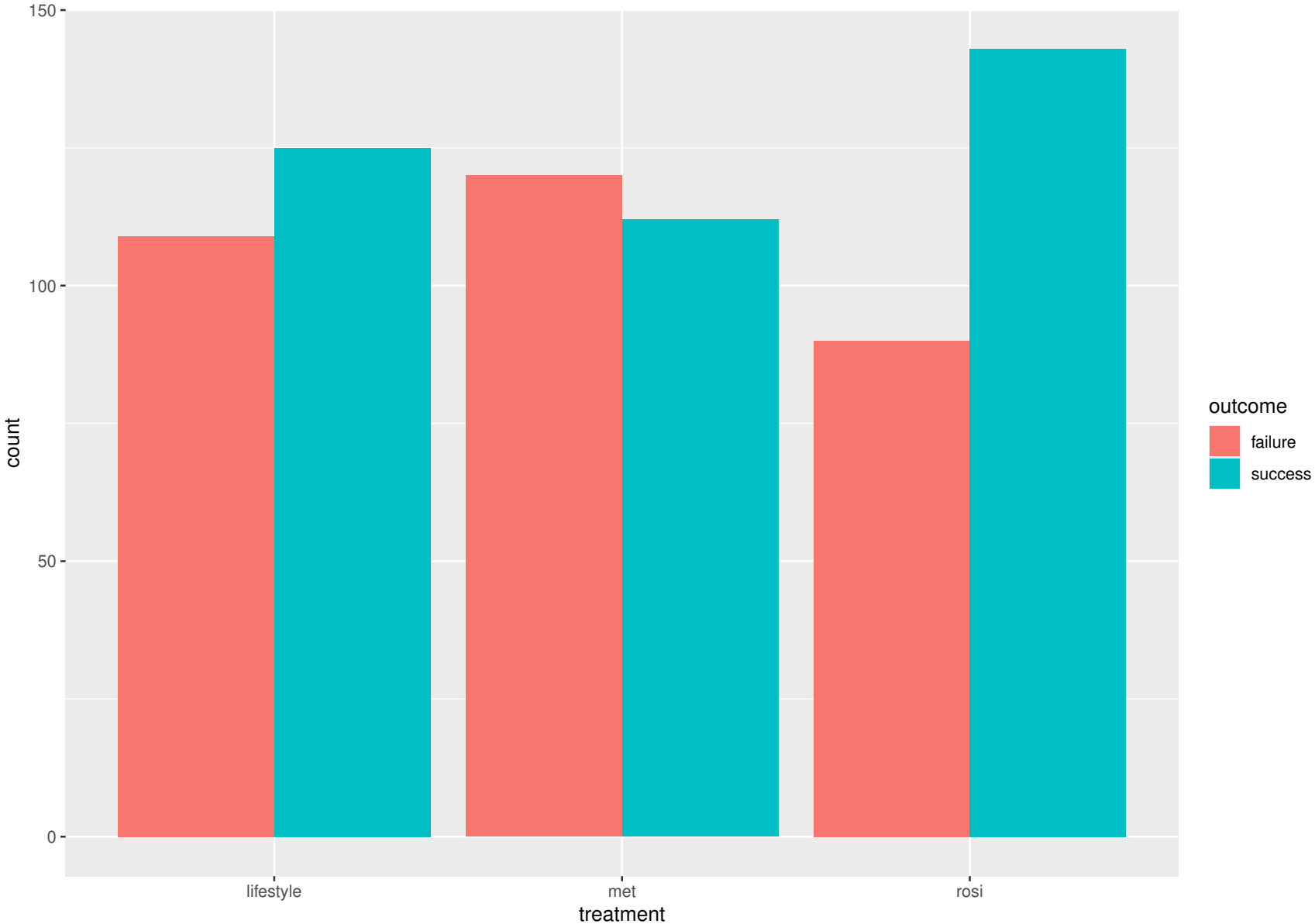
$$h_t(x) := \text{sign}(s(x) - t).$$

- **Proprietà**:
  - $s(x)$ indica log-odds $\to$ $\text{Odds}(C = 1|X = x) > e^t$ per $h_t(x) = 1$.
  - β_0 diventa $\beta_0 - t$ $\to$ $\text{Odds}(+1) := e^{-1}p(+1)/(p(-1))$
  - Per $t \to \infty$, $h_t(x) = -1$ $\to$ $FPR = 0$, $TPR = 0$
  - Per $t \to -\infty$, $h_t(x) = +1$
  - Al variare di t la performance di h_t nel piano FPR/TPR definisce la *curva ROC* (Receiver Operating Characteristic) di s (su un test/validation set).

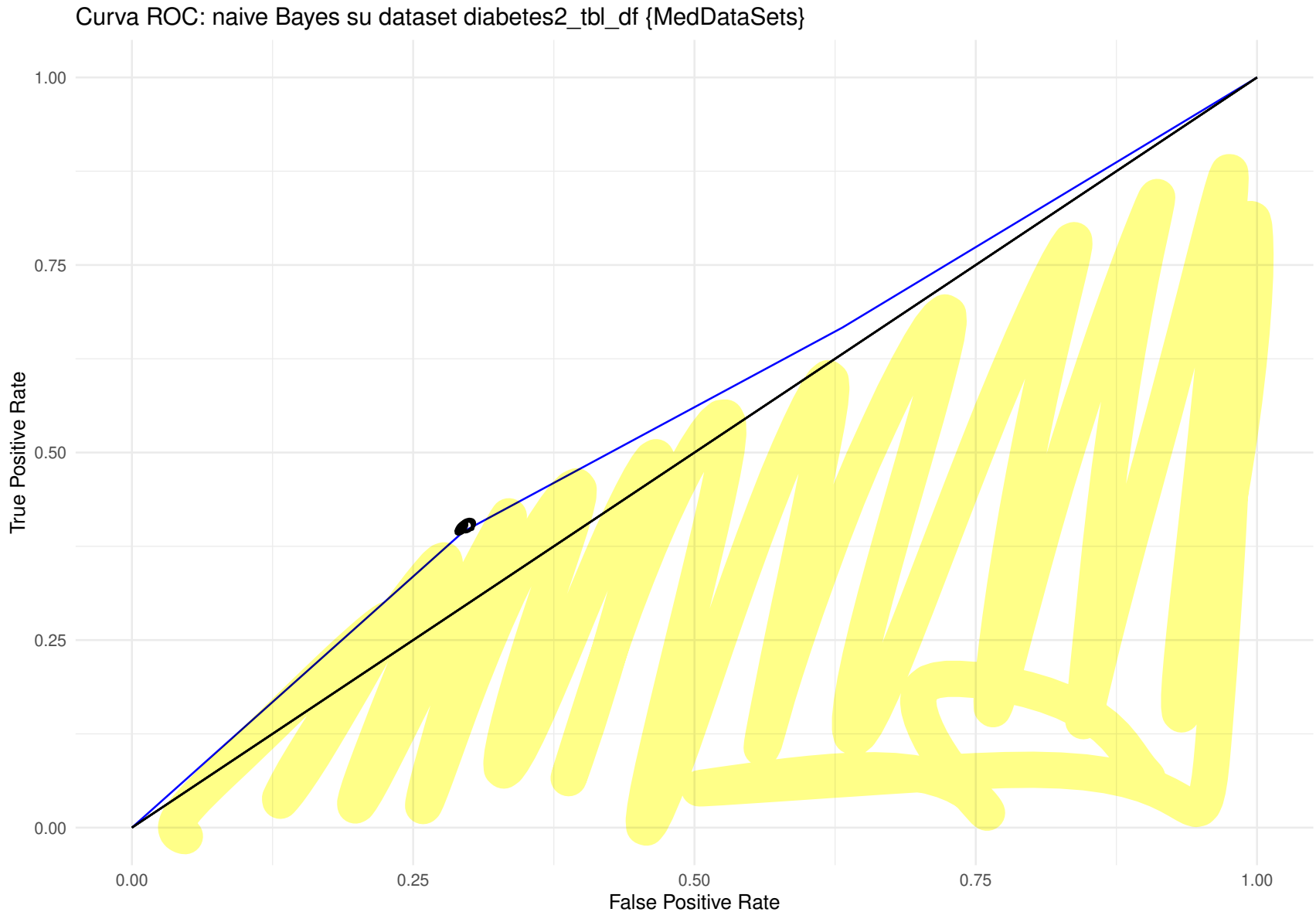
Esempio di curva ROC (dataset generato)



Dataset diabetes2_tbl_df {MedDataSets}



Dataset diabetes2_tbl_df {MedDataSets}



AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0, 1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC
- **Definizione** $AUC(s) := \int_{-\infty}^{\infty} TPR(t) dFPR(t)$

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC
- **Definizione** $AUC(s) := \int_{-\infty}^{\infty} TPR(t) dFPR(t)$
- **Interpretazione**: $AUC(s) = P(s(X_+) > s(X_-))$ dove X_+ è un *positivo reale*, X_- è un *negativo reale* e sono *indipendenti*.

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC
- **Definizione** $AUC(s) := \int_{-\infty}^{\infty} TPR(t) dFPR(t)$
- **Interpretazione**: $AUC(s) = P(s(X_+) > s(X_-))$ dove X_+ è un *positivo reale*, X_- è un *negativo reale* e sono *indipendenti*.
- **Coefficiente di Gini**: $Gini(s) = 2AUC(s) - 1$

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC
- **Definizione** $AUC(s) := \int_{-\infty}^{\infty} TPR(t) dFPR(t)$
- **Interpretazione**: $AUC(s) = P(s(X_+) > s(X_-))$ dove X_+ è un *positivo reale*, X_- è un *negativo reale* e sono *indipendenti*.
- **Coefficiente di Gini**: $Gini(s) = 2AUC(s) - 1$
- Spesso si usa $AUC(s)$ per *confrontare* score diverse (ad esempio con metodi o parametri diversi).

AUC (Area Under the Curve)

- Più la curva ROC è vicina ai lati sinistro e alto del quadrato $[0,1]^2$, migliore è la *performance* dello stimatore h_t al variare di t (quindi dello score s).
- $\leftrightarrow$ maggiore è l'area sotto la curva ROC
- **Definizione** $AUC(s) := \int_{-\infty}^{\infty} TPR(t) dFPR(t)$
- **Interpretazione**: $AUC(s) = P(s(X_+) > s(X_-))$ dove X_+ è un *positivo reale*, X_- è un *negativo reale* e sono *indipendenti*.
- **Coefficiente di Gini**: $Gini(s) = 2AUC(s) - 1$
- Spesso si usa $AUC(s)$ per *confrontare* score diverse (ad esempio con metodi o parametri diversi).
- **Attenzione!** riassumendo la curva con un numero, si perde informazione (esempio: score con curve diverse ma stessa AUC).

Gaussian Naive Bayes

Supponiamo che ciascuna caratteristica sia *reale continua*.

- *Modello*: $P(X_j = x_j | C = c) = \exp\left(-\frac{(x - m(j|c))^2}{2\sigma^2(j|c)}\right) \frac{1}{\sqrt{2\pi\sigma^2(j|c)}}$

Gaussian Naive Bayes

Supponiamo che ciascuna caratteristica sia *reale continua*.

- *Modello*: $P(X_j = x_j | C = c) = \exp\left(-\frac{(x - m(j|c))^2}{2\sigma^2(j|c)}\right) \frac{1}{\sqrt{2\pi\sigma^2(j|c)}}$
- *Parametri*:medie e varianze $(m(j|c), \sigma^2(j,c))_{j,c}$, a priori $p(c) = P(C = c) \in [0,1]^{\sharp C}$.

Gaussian Naive Bayes

Supponiamo che ciascuna caratteristica sia *reale continua*.

- *Modello*: $P(X_j = x_j | C = c) = \exp\left(-\frac{(x - m(j|c))^2}{2\sigma^2(j|c)}\right) \frac{1}{\sqrt{2\pi\sigma^2(j|c)}}$
- *Parametri*:medie e varianze $(m(j|c), \sigma^2(j,c))_{j,c}$, a priori $p(c) = P(C = c) \in [0,1]^{\sharp C}$.
- *Addestramento* (training) tramite stima (MAP/MLE): $p(c) = n(c)/n$ e

$$m(j|c) = \bar{x}_{j|c}, \quad \sigma^2(j|c) = \overline{x^2}_{j|c} - (\bar{x}_{j|c})^2,$$

Gaussian Naive Bayes

Supponiamo che ciascuna caratteristica sia *reale continua*.

- *Modello*: $P(X_j = x_j | C = c) = \exp\left(-\frac{(x - m(j|c))^2}{2\sigma^2(j|c)}\right) \frac{1}{\sqrt{2\pi\sigma^2(j|c)}}$
- *Parametri*:medie e varianze $(m(j|c), \sigma^2(j,c))_{j,c}$, a priori $p(c) = P(C = c) \in [0,1]^{\sharp C}$.
- *Addestramento* (training) tramite stima (MAP/MLE): $p(c) = n(c)/n$ e

$$m(j|c) = \bar{x}_{j|c}, \quad \sigma^2(j|c) = \bar{x^2}_{j|c} - (\bar{x}_{j|c})^2,$$

- *Classificatore*: dato $x \in F$ si calcolano:

$$P(C = c | X = x) \propto p(c) \prod_{j=1}^{d} \exp\left(-\frac{(x_j - m(j|c))^2}{2\sigma^2(j|c)}\right) \frac{1}{\sqrt{\sigma^2(j|c)}}$$

$h(x) \in \operatorname{argmax}_c P(C=c|X=x)$

Caso binario $C = \{-1, 1\}$.

- log-odds:

$$\log\left(\frac{P(C=+1|X=x)}{P(C=-1|X=x)}\right)$$

$$= \beta_0 + \sum_{j=1}^{d} \frac{(x_j - m(j|-1))^2}{2\sigma^2(j|-1)} - \frac{(x_j - m(j|+1))^2}{2\sigma^2(j|+1)}$$

Caso binario $C = \{-1, 1\}$.

- log-odds:

$$\log\left(\frac{P(C=+1|X=x)}{P(C=-1|X=x)}\right)$$
$$= \beta_0 + \sum_{j=1}^{d} \frac{(x - m(j|-1))^2}{2\sigma^2(j|-1)} - \frac{(x - m(j|+1))^2}{2\sigma^2(j|+1)}$$

- La funzione di *score* è **quadratica nelle features**:

$$s(x) := \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \alpha_j x_j^2$$

Caso binario $C = \{-1, 1\}$.

- log-odds:

$$\log\left(\frac{P(C=+1|X=x)}{P(C=-1|X=x)}\right)$$
$$= \beta_0 + \sum_{j=1}^{d} \frac{(x - m(j|-1))^2}{2\sigma^2(j|-1)} - \frac{(x - m(j|+1))^2}{2\sigma^2(j|+1)}$$

- La funzione di *score* è **quadratica nelle features**:

$$s(x) := \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \alpha_j x_j^2$$

- Se $\sigma(j|-1) = \sigma(j|+1) \rightarrow s(x)$ è **lineare** nelle features

Caso binario $C = \{-1, 1\}$.

- log-odds:

$$\log\left(\frac{P(C=+1|X=x)}{P(C=-1|X=x)}\right)$$
$$= \beta_0 + \sum_{j=1}^{d} \frac{(x - m(j|-1))^2}{2\sigma^2(j|-1)} - \frac{(x - m(j|+1))^2}{2\sigma^2(j|+1)}$$

- La funzione di *score* è **quadratica nelle features**:

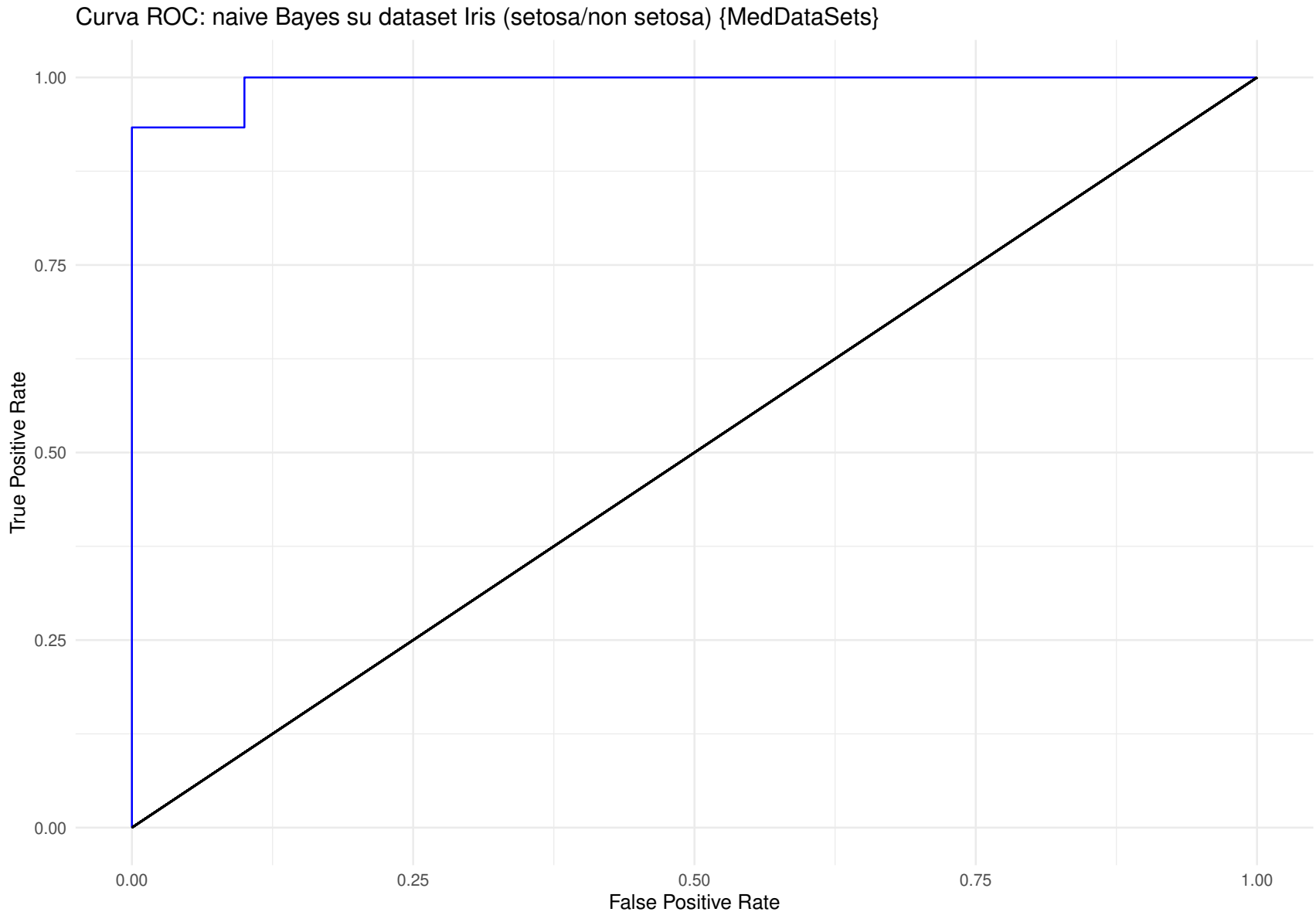
$$s(x) := \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \alpha_j x_j^2$$

- Se $\sigma(j|-1) = \sigma(j|+1) \rightarrow s(x)$ è **lineare** nelle features
- Il classificatore è $h(x) = \text{sign}(s(x)) \rightarrow$ introdurre una soglia t e porre $h_t(x) = \text{sign}(s(x) - t)$.

Esempio di Gaussian Naive Bayes (iris)

```
##              Classificati
## Reali         setosa versicolor virginica
##   setosa          15          0         0
##   versicolor       0         14         1
##   virginica        0          2        13
```

Curva ROC dataset Iris (positivo = versicolor)



LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c))}}$$

$m(c) \in \mathbb{R}^d$

$\Sigma(c) \in \mathbb{R}^{d \times d}$

↑ matrice della covarianza

LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c)}}$$

- dove $m(c)$ e $\Sigma(c)$ sono da parametri (pure le probabilità a priori $p(c)$).

LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c)}}$$

- dove $m(c)$ e $\Sigma(c)$ sono da parametri (pure le probabilità a priori $p(c)$).
- **LDA (Linear Discriminant Analysis)**:

LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c)}}$$

- dove $m(c)$ e $\Sigma(c)$ sono da parametri (pure le probabilità a priori $p(c)$).
- **LDA (Linear Discriminant Analysis)**:
  - tutte le classi con la stessa matrice di covarianza $\Sigma(c) = \Sigma$.

LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c)}}$$

- dove $m(c)$ e $\Sigma(c)$ sono da parametri (pure le probabilità a priori $p(c)$).
- **LDA (Linear Discriminant Analysis)**:
  - tutte le classi con la stessa matrice di covarianza $\Sigma(c) = \Sigma$.
- **QDA (Quadratic Discriminant Analysis)**:

LDA e QDA

- *Modello*: per features $x \in \mathbb{R}^d$, label c,

$$p(X = x|c) \propto \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{\det(\Sigma(c)}}$$

- dove $m(c)$ e $\Sigma(c)$ sono da parametri (pure le probabilità a priori $p(c)$).
- **LDA (Linear Discriminant Analysis)**:
  - tutte le classi con la stessa matrice di covarianza $\Sigma(c) = \Sigma$.
- **QDA (Quadratic Discriminant Analysis)**:
  - permette matrice di covarianza differenti per ciascuna classe.

Caso binario $C = \{-1, 1\}$:

- Modello

$p(X = x \mid c) \propto$

$$\propto p(c) \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{(2\pi)^d \det(\Sigma(c))}}$$

$p($

Caso binario $C = \{-1, 1\}$:

- Modello

polinomi di II grado nelle features

$$p(C=c|X=x) \propto p(c) \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{(2\pi)^d \det(\Sigma(c))}}$$

- In QDA: funzione di score **quadratica**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = +1|X = x)}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \sum_{j,k=1}^{d} \gamma_{j,k} x_j x_k$$

termini misti

Caso binario $C = \{-1, 1\}$:

- Modello

$$p(c) \exp\left(-\frac{1}{2}(x - m(c))^T \Sigma(c)^{-1}(x - m(c))\right) \frac{1}{\sqrt{(2\pi)^d \det(\Sigma(c))}}$$

- In QDA: funzione di score **quadratica**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = +1|X = x)}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \sum_{j,k=1}^{d} \gamma_{j,k} x_j x_k$$

- In LDA: funzione di score **lineare**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = +1|X = x)}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

Stime dei parametri (caso generale)

- Per i vettori di media (MLE):

$$m(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} x_i \quad \in \mathbb{R}^d$$

Stime dei parametri (caso generale)

- Per i vettori di media (MLE):

$$m(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} x_i$$

- Per la matrice di covarianza (MLE):

Stime dei parametri (caso generale)

- Per i vettori di media (MLE):

$$m(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} x_i$$

- Per la matrice di covarianza (MLE):
  - Caso LDA:

$$\Sigma_{MLE} = \frac{1}{n} \sum_{c} \sum_{i:\ell_i=c} (x_i - m(c))(x_i - m(c))^T$$

Stime dei parametri (caso generale)

- Per i vettori di media (MLE):

$$m(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} x_i$$

- Per la matrice di covarianza (MLE):
  - Caso LDA:

$$\Sigma_{MLE} = \frac{1}{n} \sum_{c} \sum_{i:\ell_i=c} (x_i - m(c))(x_i - m(c))^T$$

  - Caso QDA:

$$\Sigma(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} (x_i - m(c))(x_i - m(c))^T$$

Stime dei parametri (caso generale)

- Per i vettori di media (MLE):

$$m(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} x_i$$

- . Per la matrice di covarianza (MLE):
  - Caso LDA:

$$\Sigma_{MLE} = \frac{1}{n} \sum_{c} \sum_{i:\ell_i=c} (x_i - m(c))(x_i - m(c))^T$$

  - Caso QDA:

$$\Sigma(c) = \frac{1}{n(c)} \sum_{i:\ell_i=c} (x_i - m(c))(x_i - m(c))^T$$

- **Attenzione:** ci sono anche altri criteri per la stima (Fisher)

Regressione Logistica

Consideriamo prima il caso *binario* $C = \{-1, +1\}$.

- Modello **discriminativo** con score **lineare**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = -1|X = x)}\right)$$

$$:= \beta_0 + \sum_{j=1}^{d} \beta_j x_j = \beta x$$

Regressione Logistica

Consideriamo prima il caso *binario* $C = \{-1, +1\}$.

- Modello **discriminativo** con score **lineare**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = -1|X = x)}\right)$$
$$:= \beta_0 + \sum_{j=1}^{d} \beta_j x_j = \beta x$$

- Probabilità

Regressione Logistica

Consideriamo prima il caso *binario* $C = \{-1, +1\}$.

- Modello **discriminativo** con score **lineare**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = -1|X = x)}\right)$$
$$:= \beta_0 + \sum_{j=1}^{d} \beta_j x_j = \beta x$$

- Probabilità
  - $P(C = +1|X = x) = \frac{\exp(\beta x)}{1+\exp(\beta x)}$

Regressione Logistica

Consideriamo prima il caso *binario* $C = \{-1, +1\}$.

- Modello **discriminativo** con score **lineare**

$$s(x) = \log\left(\frac{P(C = +1|X = x)}{P(C = -1|X = x)}\right)$$
$$:= \beta_0 + \sum_{j=1}^{d} \beta_j x_j = \beta x$$

- Probabilità
  - $P(C = +1|X = x) = \frac{\exp(\beta x)}{1+\exp(\beta x)}$
  - $P(C = -1|X = x) = \frac{1}{1+\exp(\beta x)}$

Verosimiglianza regressione logistica

- Per indipendenza (delle osservazioni):

$$\mathcal{L}((\beta_j)_{j=0}^d; (x_i, \ell_i)_{i=1,\dots,n}) = \prod_{i=1}^n P(X = x_i | C = \ell_i)$$

$$= \frac{\exp\left(\beta \sum_{i:\ell_i=1} x_i\right)}{\prod_{i=1}^n (1 + \exp(\beta x_i))}$$

Stimo i parametri β tramite MLE ↑

Verosimiglianza regressione logistica

- Per indipendenza:

$$\mathcal{L}((\beta_j)_{j=0}^d; (x_i, \ell_i)_{i=1,\dots,n}) = \prod_{i=1}^n P(X = x_i | C = \ell_i)$$
$$= \frac{\exp\left(\beta \sum_{i:\ell_i=1} x\right)}{\prod_{i=1}^n (1 + \exp(\beta x_i))}$$

- Per determinare β_{MLE} si usano metodi di **ottimizzazione numerica** (la funzione $\mathcal{L}$ è convessa in β)

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$
   - Per $x \in F$, si confrontano le probabilità (o gli scores) di ciascuna label:

$$h(x) \in \arg\max_{c} \beta(c)x$$

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$
   - Per $x \in F$, si confrontano le probabilità (o gli scores) di ciascuna label:
   $$h(x) \in \arg\max_c \beta(c)x$$

2. **One-vs-One**:

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$
   - Per $x \in F$, si confrontano le probabilità (o gli scores) di ciascuna label:
   $$h(x) \in \arg\max_c \beta(c)x$$

2. **One-vs-One**:

- Per ogni coppia di labels c, ℓ si addestra un classificatore binario $h_{c,\ell}$ che distingue c da ℓ.

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$
   - Per $x \in F$, si confrontano le probabilità (o gli scores) di ciascuna label:

$$h(x) \in \arg\max_{c} \beta(c)x$$

2. **One-vs-One**:

- Per ogni coppia di labels c, ℓ si addestra un classificatore binario $h_{c,\ell}$ che distingue c da ℓ.
- Per classificare $x \in F$, ciascun classificatore **esprime un voto** $h_{c,\ell}(x) \in \{c, \ell\}$

Regressione Logistica: Caso Multi-Classe

Si possono utilizzare diverse strategie *euristiche*:

1. **One-vs-All (One-vs-Rest)**:
   - Per ogni label c, si addestra un classificatore binario h_c che distingue c da tutte le altre con parametri $(\beta(c)_j)_{j=0}^{d}$
   - Per $x \in F$, si confrontano le probabilità (o gli scores) di ciascuna label:

$$h(x) \in \arg\max_c \beta(c)x$$

2. **One-vs-One**:

- Per ogni coppia di labels c, ℓ si addestra un classificatore binario $h_{c,\ell}$ che distingue c da ℓ.
- Per classificare $x \in F$, ciascun classificatore **esprime un voto** $h_{c,\ell}(x) \in \{c, \ell\}$
- si pone $h(x) :=$ la label che riceve più voti.

Regressione softmax

- Generalizzare la regressione logistica a multi-classe usando la fuzione *softmax* per convertire i punteggi in probabilità.

Regressione softmax

- Generalizzare la regressione logistica a multi-classe usando la fuzione *softmax* per convertire i punteggi in probabilità.
- Per ogni classe c, si introducono parametri $\beta(c)$.

Regressione softmax

- Generalizzare la regressione logistica a multi-classe usando la fuzione *softmax* per convertire i punteggi in probabilità.
- Per ogni classe c, si introducono parametri $\beta(c)$.
- La funzione *softmax* è definita come:

$$P(C = c|X = x) = \frac{\exp(\beta(c)x)}{\sum_{\ell} \exp(\beta(\ell)x)}$$

Regressione softmax

- Generalizzare la regressione logistica a multi-classe usando la fuzione *softmax* per convertire i punteggi in probabilità.
- Per ogni classe c, si introducono parametri $\beta(c)$.
- La funzione *softmax* è definita come:

$$P(C = c|X = x) = \frac{\exp(\beta(c)x)}{\sum_{\ell} \exp(\beta(\ell)x)}$$

- Si ottimizzano i parametri tramite MLE (numericamente)

Regressione softmax

- Generalizzare la regressione logistica a multi-classe usando la fuzione *softmax* per convertire i punteggi in probabilità.
- Per ogni classe c, si introducono parametri $\beta(c)$.
- La funzione *softmax* è definita come:

$$P(C = c|X = x) = \frac{\exp(\beta(c)x)}{\sum_{\ell} \exp(\beta(\ell)x)}$$

- Si ottimizzano i parametri tramite MLE (numericamente)
- permette di trattare direttamente più classi senza la necessità di addestrare classificatori separati.