Naive Bayes Classifier

Jérémie Cabessa

Laboratoire DAVID, UVSQ

Introduction

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**
  La variable d'output (réponse) est **quantitative**.
- **Méthodes de classification**
  La variable d'output (réponse) est **qualitative**.

Introduction

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**
  La variable d'output (réponse) est **quantitative**.
- **Méthodes de classification**
  La variable d'output (réponse) est **qualitative**.

Introduction

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**
  La variable d'output (réponse) est **quantitative**.
- **Méthodes de classification**
  La variable d'output (réponse) est **qualitative**.

Introduction

- Un **classifieur de Bayes naïf (naive Bayes classifier)** est une méthode de classification basée sur le *théorème de Bayes*.
- Un naive Bayes classifier possède un petit nombre de paramètres à estimer: linéaire par rapport au nombre de variables.
- L'apprentissage de ces paramètres admet une solution analytique (closed-form solution) calculable en temps linéaire.

Introduction

- Un **classifieur de Bayes naïf (naive Bayes classifier)** est une méthode de classification basée sur le *théorème de Bayes*.
- Un naive Bayes classifier possède un petit nombre de paramètres à estimer: linéaire par rapport au nombre de variables.
- L'apprentissage de ces paramètres admet une solution analytique (closed-form solution) calculable en temps linéaire.

Introduction

- Un **classifieur de Bayes naïf (naive Bayes classifier)** est une méthode de classification basée sur le *théorème de Bayes*.
- Un naive Bayes classifier possède un petit nombre de paramètres à estimer: linéaire par rapport au nombre de variables.
- L'apprentissage de ces paramètres admet une solution analytique (closed-form solution) calculable en temps linéaire.

Probabilités conditionnelles

- Soit p une mesure de probabilité et A et B deux évènements de probabilité non nulle.
- La *probabilité conditionnelle* de A sachant B est

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

- $p(A \mid B)$ représente *la probabilité que l'évènement A advienne sachant que l'évènement B a eu lieu.*

Probabilités conditionnelles

- Soit p une mesure de probabilité et A et B deux évènements de probabilité non nulle.
- La *probabilité conditionnelle* de A sachant B est

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

- $p(A \mid B)$ représente *la probabilité que l'évènement A advienne sachant que l'évènement B a eu lieu.*

Probabilités conditionnelles

- Soit p une mesure de probabilité et A et B deux évènements de probabilité non nulle.
- La *probabilité conditionnelle* de A sachant B est

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

- $p(A \mid B)$ représente *la probabilité que l'évènement A advienne sachant que l'évènement B a eu lieu*.

Théorème de Bayes

- La formule de la *probabilité conditionnelle* implique les relations suivantes:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \quad \Rightarrow \quad p(A \cap B) = p(A \mid B)\, p(B)$$

$$p(B \mid A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} \quad \Rightarrow \quad p(A \cap B) = p(B \mid A)\, p(A)$$

- Les équations ci-dessus impliquent le théorème de Bayes:

$$p(A \mid B)\, p(B) = p(B \mid A)\, p(A)$$

$$\Leftrightarrow$$

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Théorème de Bayes

- La formule de la *probabilité conditionnelle* implique les relations suivantes:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \quad \Rightarrow \quad p(A \cap B) = p(A \mid B)\, p(B)$$

$$p(B \mid A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} \quad \Rightarrow \quad p(A \cap B) = p(B \mid A)\, p(A)$$

Théorème de Bayes

- La formule de la *probabilité conditionnelle* implique les relations suivantes:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \quad \Rightarrow \quad p(A \cap B) = p(A \mid B)\, p(B)$$

$$p(B \mid A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} \quad \Rightarrow \quad p(A \cap B) = p(B \mid A)\, p(A)$$

- Les equations ci-dessus impliquent le théorème de Bayes:

$$p(A \mid B)\, p(B) \;=\; p(B \mid A)\, p(A)$$

ssi

$$p(A \mid B) \;=\; \frac{p(B \mid A)\, p(A)}{p(B)}$$

Théorème de Bayes

- La formule de la *probabilité conditionnelle* implique les relations suivantes:

$$
\begin{aligned}
p(A \mid B) &= \frac{p(A \cap B)}{p(B)} && \Rightarrow \quad p(A \cap B) = p(A \mid B)\, p(B) \\
p(B \mid A) &= \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} && \Rightarrow \quad p(A \cap B) = p(B \mid A)\, p(A)
\end{aligned}
$$

- Les equations ci-dessus impliquent le théorème de Bayes:

$$p(A \mid B)\, p(B) = p(B \mid A)\, p(A)$$

ssi

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Théorème de Bayes

- La formule de la *probabilité conditionnelle* implique les relations suivantes:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \quad \Rightarrow \quad p(A \cap B) = p(A \mid B)\, p(B)$$

$$p(B \mid A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} \quad \Rightarrow \quad p(A \cap B) = p(B \mid A)\, p(A)$$

- Les equations ci-dessus impliquent le **théorème de Bayes**:

$$p(A \mid B)\, p(B) = p(B \mid A)\, p(A)$$

ssi

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Théorème de Bayes

- Le théorème de Bayes s'utilise lorsqu'on désire calculer $p(A \mid B)$ mais que cette quantité est difficile à estimer à partir des data.

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

- Il est alors possible de calculer $p(A \mid B)$ à partir de $p(B \mid A)$, qui et potentiellement plus facile à estimer à partir des data.
- En résumé, la probabilité conditionnelle $p(A \mid B)$ peut s'exprimer à partir de sa probabilité conditionnelle "inverse" $p(B \mid A)$.

Théorème de Bayes

- Le théorème de Bayes s'utilise lorsqu'on désire calculer $p(A \mid B)$ mais que cette quantité est difficile à estimer à partir des data.

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

- Il est alors possible de calculer $p(A \mid B)$ à partir de $p(B \mid A)$, qui et potentiellement plus facile à estimer à partir des data.
- En résumé, la probabilité conditionnelle $p(A \mid B)$ peut s'exprimer à partir de sa probabilité conditionnelle "inverse" $p(B \mid A)$.

Théorème de Bayes

- Le théorème de Bayes s'utilise lorsqu'on désire calculer $p(A \mid B)$ mais que cette quantité est difficile à estimer à partir des data.

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

- Il est alors possible de calculer $p(A \mid B)$ à partir de $p(B \mid A)$, qui et potentiellement plus facile à estimer à partir des data.
- En résumé, la probabilité conditionnelle $p(A \mid B)$ peut s'exprimer à partir de sa probabilité conditionnelle “inverse” $p(B \mid A)$.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.

- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.

- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.
- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.

- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.

- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Basé sur une étude de Kahneman & Tversky:

  Steve est d'un tempérament doux, timide, plutôt introverti. Bien qu'il soit très aimable, il semble montrer peu d'attention envers les autres personnes. Il a tendance à être très ordonnée et montre un intérêt marqué pour tout ce qui est de l'ordre du détail.
- Laquelle de ces affirmations semble la plus plausible?

(A) Steve est libraire.

(B) Steve est agriculteur.

Théorème de Bayes

- Si on laisse de côté la question des *préjugés* ou des *stéréotypes* que l'on peut avoir sur différentes professions, la plupart des gens répondent que:
- Steve a bien plus de chance d'être libraire qu'agriculteur.
- Mais cette réponse est *irrationnelle*... Pourquoi?

Théorème de Bayes

- Si on laisse de côté la question des *préjugés* ou des *stéréotypes* que l'on peut avoir sur différentes professions, la plupart des gens répondent que:
- Steve a bien plus de chance d'être libraire qu'agriculteur.
- Mais cette réponse est *irrationnelle*... Pourquoi?

Théorème de Bayes

- Si on laisse de côté la question des *préjugés* ou des *stéréotypes* que l'on peut avoir sur différentes professions, la plupart des gens répondent que:
- Steve a bien plus de chance d'être libraire qu'agriculteur.
- Mais cette réponse est *irrationnelle*... Pourquoi?

Théorème de Bayes

- Les gens oublient totalement de prendre en compte la proportion de libraire libraires et d'agriculteurs dans la population!
- Aux USA, il semble qu'il y ait bien plus d'agriculteurs que de libraires!
- Supposons que la population compte 20 fois plus d'agriculteurs que de libraires: même si Steve semble montrer des "traits" de libraires, il reste assez peu probable qu'il soit effectivement libraire...

Théorème de Bayes

- Les gens oublient totalement de prendre en compte la proportion de libraire libraires et d'agriculteurs dans la population!
- Aux USA, il semble qu'il y ait bien plus d'agriculteurs que de libraires!
- Supposons que la population compte 20 fois plus d'agriculteurs que de libraires: même si Steve semble montrer des "traits" de libraires, il reste assez peu probable qu'il soit effectivement libraire...

Théorème de Bayes

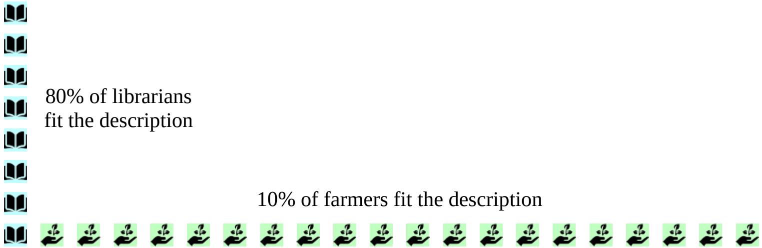
- Les gens oublient totalement de prendre en compte la proportion de libraire libraires et d'agriculteurs dans la population!
- Aux USA, il semble qu'il y ait bien plus d'agriculteurs que de libraires!
- Supposons que la population compte 20 fois plus d'agriculteurs que de libraires: même si Steve semble montrer des "traits" de libraires, il reste assez peu probable qu'il soit effectivement libraire...

Théorème de Bayes

- On a 10 libraires et 200 agriculteurs. De plus, 80% des libraires et 10% des agriculteurs et correspondent à la description.
- $p(libraire \mid description) = \frac{8}{8+20} \approx 28.57\%$

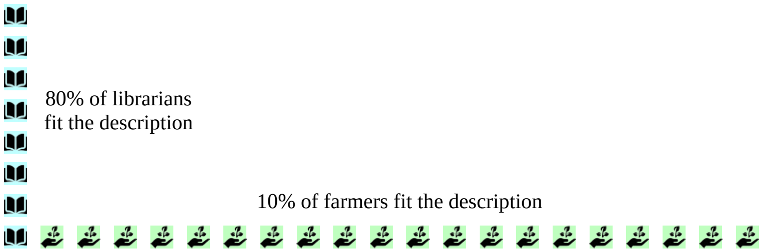
Théorème de Bayes

- On a 10 libraires et 200 agriculteurs. De plus, 80% des libraires et 10% des agriculteurs et correspondent à la description.
- $p(\textit{libraire} \mid \textit{description}) = \frac{8}{8+20} \approx 28.57\%$



Théorème de Bayes

- On a 10 libraires et 200 agriculteurs. De plus, 80% des libraires et 10% des agriculteurs et correspondent à la description.
- $p(libraire \mid description) = \frac{8}{8+20} \approx 28.57\%$



Théorème de Bayes

- **Hypothèse** H**:** Steve est libraire.
- **Évidence** E**:** Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Hypothèse** H: Steve est libraire.
- **Évidence** E: Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Hypothèse** H: Steve est libraire.
- **Évidence** E: Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Hypothèse** H**:** Steve est libraire.
- **Évidence** E**:** Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Hypothèse H:** Steve est libraire.
- **Évidence E:** Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Hypothèse** H**:** Steve est libraire.
- **Évidence** E**:** Steve est d'un tempérament doux, timide, ...
- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Likelihood:** Probabilité de l'évidence étant donné que l'hypothèse est vraie $p(E \mid H) = 80\%$.
- **(Likelihood bis):** Probabilité de l'évidence étant donné que l'hypothèse est fausse $p(E \mid \neg H) = 10\%$.
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).

Théorème de Bayes

- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).
- Notre croyance que Steve est libraire est passée de 4.76% (croyance a priori) à 28.57% (croyance a posteriori).
- La différence entre le *prior* et le *posterior* est appelée **belief updating** ou **belief revision** (**révision des croyances**).

Théorème de Bayes

- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).
- Notre croyance que Steve est libraire est passée de 4.76% (croyance a priori) à 28.57% (croyance a posteriori).
- La différence entre le *prior* et le *posterior* est appelée **belief updating** ou **belief revision (révision des croyances)**.

Théorème de Bayes

- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).
- Notre croyance que Steve est libraire est passée de 4.76% (croyance a priori) à 28.57% (croyance a posteriori).
- La différence entre le *prior* et le *posterior* est appelée **belief updating** ou **belief revision** (révision des croyances).

Théorème de Bayes

- **Prior:** Probabilité de l'hypothèse avant de recevoir une évidence: $p(H) = \frac{10}{210} = \frac{1}{21} \approx 4.76\%$
- **Posterior:** Probabilité de l'hypothèse étant donné l'évidence $p(H \mid E) = \frac{8}{8+20} \approx 28.57\%$ (ce que l'on cherche).
- Notre croyance que Steve est libraire est passée de 4.76% (croyance a priori) à 28.57% (croyance a posteriori).
- La différence entre le *prior* et le *posterior* est appelée **belief updating** ou **belief revision (révision des croyances)**.

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\frac{\cancel{10}}{\cancel{210}} \frac{80}{\cancel{100}}}{\frac{\cancel{10}}{\cancel{210}} \frac{80}{\cancel{100}} + \frac{200}{\cancel{210}} \frac{\cancel{10}}{\cancel{100}}} \\
&= \frac{p(H)\, p(E \mid H)}{p(H)\, p(E \mid H) + p(\neg H)\, p(E \mid \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le théorème de Bayes

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\cancel{\frac{10}{210}} \, \frac{80}{\cancel{100}}}{\cancel{\frac{10}{210}} \, \frac{80}{\cancel{100}} + \frac{200}{\cancel{210}} \, \frac{\cancel{10}}{\cancel{100}}} \\
&= \frac{p(H)\, p(E \mid H)}{p(H)\, p(E \mid H) + p(\neg H)\, p(E \mid \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le théorème de Bayes

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\frac{\cancel{10}}{\cancel{210}}\,\frac{80}{\cancel{100}}}{\frac{\cancel{10}}{\cancel{210}}\,\frac{80}{\cancel{100}} + \frac{200}{\cancel{210}}\,\frac{\cancel{10}}{\cancel{100}}} \\
&= \frac{p(H)\,p(E \mid H)}{p(H)\,p(E \mid H) + p(\neg H)\,p(E \mid \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le théorème de Bayes

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\cancel{\frac{10}{210}}\,\cancel{\frac{80}{100}}}{\cancel{\frac{10}{210}}\,\cancel{\frac{80}{100}} + \cancel{\frac{200}{210}}\,\cancel{\frac{10}{100}}} \\
&= \frac{p(H)\,p(E \mid H)}{p(H)\,p(E \mid H) + p(\neg H)\,p(E \mid \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le théorème de Bayes.

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\frac{\cancel{10}}{\cancel{210}}\,\frac{80}{\cancel{100}}}{\frac{\cancel{10}}{\cancel{210}}\,\frac{80}{\cancel{100}} + \frac{200}{\cancel{210}}\,\frac{\cancel{10}}{\cancel{100}}} \\
&= \frac{p(H)\,p(E \mid H)}{p(H)\,p(E \mid H) + p(\neg H)\,p(E \mid \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\,p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le théorème de Bayes.

Théorème de Bayes

- On a donc

$$
\begin{aligned}
p(H \mid E) &= \frac{8}{8+20} \\
&= \frac{\frac{\cancel{10}}{\cancel{210}} \frac{80}{\cancel{100}}}{\frac{\cancel{10}}{\cancel{210}} \frac{80}{\cancel{100}} + \frac{200}{\cancel{210}} \frac{\cancel{10}}{\cancel{100}}} \\
&= \frac{p(H)\, p(E \mid H)}{p(H)\, p(E \mid H) + p(\neg H)\, p(E \mid \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E \cap H) + p(E \cap \neg H)} \\
&= \frac{p(H)\, p(E \mid H)}{p(E)}
\end{aligned}
$$

- On retrouve alors le **théorème de Bayes**.

Modèle probabiliste

- Soient $\boldsymbol{X} = (X_1, \ldots, X_P)$ des variables explicatives et Y une variable réponse qualitative à valeurs dans $C = \{c_1, \ldots, c_K\}$.
- Soit un train set

$$S = \{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^P \times C : i = 1, \ldots, N\}.$$

Modèle probabiliste

- Soient $\boldsymbol{X} = (X_1, \ldots, X_P)$ des variables explicatives et Y une variable réponse qualitative à valeurs dans $C = \{c_1, \ldots, c_K\}$.
- Soit un train set

$$S = \big\{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^P \times C : i = 1, \ldots, N\big\}.$$

Modèle probabiliste

- Le **modèle probabiliste** pour un classifieur consiste à calculer la probabilité conditionnelle, étant donné un point $\boldsymbol{x}$, d'appartenir à chacune des classes c_k, i.e.

$$p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x}) \text{ pour tout } k = 1, \ldots, K$$

- Ensuite, on associe x à la classe $\hat{c}$ dont la probabilité conditionnelle est maximale, i.e.,

$$\hat{c} = \arg\max_{c_k \in C} p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$$

Modèle probabiliste

- Le **modèle probabiliste** pour un classifieur consiste à calculer la probabilité conditionnelle, étant donné un point $\boldsymbol{x}$, d'appartenir à chacune des classes c_k, i.e.

$$p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x}) \text{ pour tout } k = 1, \ldots, K$$

- Ensuite, on associe $\boldsymbol{x}$ à la classe $\hat{c}$ dont la probabilité conditionnelle est maximale, i.e.,

$$\hat{c} = \arg\max_{c_k \in C} p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$p(c_k, x_1, \dots, x_P) = p(x_1, \dots, x_P, c_k)$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k)\, p(x_2, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k)\, p(x_3, \ldots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad\; \cdots \\
&\quad\; p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k)\, p(x_2, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k)\, p(x_3, \ldots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad\; \cdots \\
&\quad\; p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k)\, p(x_2, \ldots, x_P, c_k) \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k)\, p(x_3, \ldots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad\; \cdots \\
&\quad\; p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \dots, x_P) &= p(x_1, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k)\, p(x_2, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \dots, x_P, c_k)\, p(x_3, \dots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\; p(x_2 \mid x_3, \dots, x_P, c_k) \\
&\quad\; \cdots \\
&\quad\; p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \dots, x_P) &= p(x_1, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k)\, p(x_2, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\ p(x_2 \mid x_3, \dots, x_P, c_k)\, p(x_3, \dots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\ p(x_2 \mid x_3, \dots, x_P, c_k) \\
&\quad\ \cdots \\
&\quad\ p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Modèle probabiliste

- Pour un point $\boldsymbol{x}$ on abrège $p(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x})$ par $p(c_k \mid \boldsymbol{x})$.
- Une application répétée de la règles des probabilités conditionnelles donne:

$$
\begin{aligned}
p(c_k, x_1, \dots, x_P) &= p(x_1, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k)\, p(x_2, \dots, x_P, c_k) \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\ p(x_2 \mid x_3, \dots, x_P, c_k)\, p(x_3, \dots, x_P, c_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \dots, x_P, c_k) \\
&\quad\ p(x_2 \mid x_3, \dots, x_P, c_k) \\
&\quad\ \dots \\
&\quad\ p(x_{P-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k)
\end{aligned}
$$

Hypothèse naïve

- On introduit l'hypothèse naïve d'*indépendance conditionnelle*: chaque X_i est indépendant des autres caractéristiques X_j, conditionnellement à Y, i.e.

$$p(x_j \mid x_{j+1}, \dots, x_P, c_k) = p(x_j \mid c_k)$$

- **Exemple:** Classification de fruits à partir de divers attributs (couleur, forme, etc.). L'hypothèse naïve dit:

$$p(X_1 = rouge \mid X_2 = rond, Y = pomme)$$
$$= p(X_1 = rouge \mid Y = pomme)$$

Hypothèse naïve

- On introduit l'hypothèse naïve d'*indépendance conditionnelle*: chaque X_i est indépendant des autres caractéristiques X_j, conditionnellement à Y, i.e.

$$p(x_j \mid x_{j+1}, \ldots, x_P, c_k) = p(x_j \mid c_k)$$

- **Exemple:** Classification de fruits à partir de divers attributs (couleur, forme, etc.). L'hypothèse naïve dit:

$$p(X_1 = rouge \mid X_2 = rond, Y = pomme) \\ = p(X_1 = rouge \mid Y = pomme)$$

Hypothèse naïve

- En appliquant l'hypothèse naïve, on a:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad \ldots \\
&\quad p(x_{n-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k) \\
&= p(x_1 \mid c_k)\, p(x_2 \mid c_k)\, p(x_3 \mid c_k) \cdots p(c_k) \\
&= p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \qquad (1)
\end{aligned}
$$

Hypothèse naïve

- En appliquant l'hypothèse naïve, on a:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad \ldots \\
&\quad p(x_{n-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k) \\
&= p(x_1 \mid c_k)\, p(x_2 \mid c_k)\, p(x_3 \mid c_k) \cdots p(c_k) \\
&= p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \qquad (1)
\end{aligned}
$$

Hypothèse naïve

- En appliquant l'hypothèse naïve, on a:

$$\begin{aligned} p(c_k, x_1, \ldots, x_P) &= p(x_1 \mid x_2, \ldots, x_P, c_k) \\ &\quad p(x_2 \mid x_3, \ldots, x_P, c_k) \\ &\quad \ldots \\ &\quad p(x_{n-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k) \\ &= p(x_1 \mid c_k)\, p(x_2 \mid c_k)\, p(x_3 \mid c_k) \cdots p(c_k) \\ &= p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \qquad (1) \end{aligned}$$

Hypothèse naïve

- En appliquant l'hypothèse naïve, on a:

$$
\begin{aligned}
p(c_k, x_1, \ldots, x_P) &= p(x_1 \mid x_2, \ldots, x_P, c_k) \\
&\quad p(x_2 \mid x_3, \ldots, x_P, c_k) \\
&\quad \ldots \\
&\quad p(x_{n-1} \mid x_P, c_k)\, p(x_P \mid c_k)\, p(c_k) \\
&= p(x_1 \mid c_k)\, p(x_2 \mid c_k)\, p(x_3 \mid c_k) \cdots p(c_k) \\
&= p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \qquad (1)
\end{aligned}
$$

Naive Bayes Classifier

- Par le théorème de Bayes et l'équation (1), on a finalement:

$$p(c_k \mid x_1, \ldots, x_P) = \frac{p(c_k, x_1, \ldots, x_P)}{p(x_1, \ldots, x_P)} = \frac{p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)}{p(x_1, \ldots, x_P)}$$

- En résumé, la formule d'un **naive Bayes classifier** est:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Ainsi, la **prédiction** $\hat{c}$ du classifieur est donnée par

$$\hat{c} = \arg\max_{c_k \in C} p(c_k \mid \boldsymbol{x}) = \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

Naive Bayes Classifier

- Par le théorème de Bayes et l'équation (1), on a finalement:

$$p(c_k \mid x_1, \dots, x_P) = \frac{p(c_k, x_1, \dots, x_P)}{p(x_1, \dots, x_P)} = \frac{p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)}{p(x_1, \dots, x_P)}$$

- En résumé, la formule d'un **naive Bayes classifier** est:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Ainsi, la **prédiction** $\hat{c}$ du classifieur est donnée par

$$\hat{c} = \arg\max_{c_k \in C} p(c_k \mid \boldsymbol{x}) = \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

Naive Bayes Classifier

- Par le théorème de Bayes et l'équation (1), on a finalement:

$$p(c_k \mid x_1, \ldots, x_P) = \frac{p(c_k, x_1, \ldots, x_P)}{p(x_1, \ldots, x_P)} = \frac{p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)}{p(x_1, \ldots, x_P)}$$

- En résumé, la formule d'un **naive Bayes classifier** est:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Ainsi, la **prédiction** $\hat{c}$ du classifieur est donnée par

$$\hat{c} = \arg\max_{c_k \in C} p(c_k \mid \boldsymbol{x}) = \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

Estimation des paramètres

- On a donc:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \;\propto\; p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Grâce au théorème de Bayes et à notre hypothèse naïve, on a pu exprimer la probabilité conditionnelle $p(c_k \mid \boldsymbol{x})$ à partir de $p(c_k)$ et des probabilités conditionnelles “inverses” $p(x_j \mid c_k)$.
- Mais que valent $p(c_k)$ et $p(x_j \mid c_k)$? Comment estimer $p(c_k)$ et $p(x_j \mid c_k)$ à partir des data?

Estimation des paramètres

- On a donc:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \; \propto \; p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Grâce au théorème de Bayes et à notre hypothèse naïve, on a pu exprimer la probabilité conditionnelle $p(c_k \mid \boldsymbol{x})$ à partir de $p(c_k)$ et des probabilités conditionnelles "inverses" $p(x_j \mid c_k)$.
- Mais que valent $p(c_k)$ et $p(x_j \mid c_k)$? Comment estimer $p(c_k)$ et $p(x_j \mid c_k)$ à partir des data?

Estimation des paramètres

- On a donc:

$$p(c_k \mid \boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \;\propto\; p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

- Grâce au théorème de Bayes et à notre hypothèse naïve, on a pu exprimer la probabilité conditionnelle $p(c_k \mid \boldsymbol{x})$ à partir de $p(c_k)$ et des probabilités conditionnelles "inverses" $p(x_j \mid c_k)$.
- Mais que valent $p(c_k)$ et $p(x_j \mid c_k)$? Comment estimer $p(c_k)$ et $p(x_j \mid c_k)$ à partir des data?

Estimation des priors

- **Estimation des "class priors":** l'estimation des $p(Y = c_k) = p(c_k)$ pour $k = 1, \ldots, K$ à partir des data est simple.
- Soit on suppose que toutes les K classes $c_1, \ldots, c_K$ sont équiprobables, auquel cas on a:

$$p(c_k) = \frac{1}{K} \text{ , pour tout } k = 1, \ldots, K.$$

- Ou alors on estime $p(c_k)$ comme la proportion d'éléments du train set S qui sont de la classe c_k, i.e.

$$p(c_k) = \frac{|S_k|}{N} \text{ , pour tout } k = 1, \ldots, K.$$

où S_k est le sous-dataset de S formé des éléments de classe c_k.

Estimation des priors

- **Estimation des "class priors":** l'estimation des $p(Y = c_k) = p(c_k)$ pour $k = 1, \ldots, K$ à partir des data est simple.
- Soit on suppose que toutes les K classes $c_1, \ldots, c_K$ sont équiprobables, auquel cas on a:

$$p(c_k) = \frac{1}{K} \text{ , pour tout } k = 1, \ldots, K.$$

- Ou alors on estime $p(c_k)$ comme la proportion d'éléments du train set S qui sont de la classe c_k, i.e.

$$p(c_k) = \frac{|S_k|}{N} \text{ , pour tout } k = 1, \ldots, K.$$

où S_k est le sous-dataset de S formé des éléments de classe c_k.

Estimation des priors

- **Estimation des "class priors":** l'estimation des $p(Y = c_k) = p(c_k)$ pour $k = 1, \dots, K$ à partir des data est simple.
- Soit on suppose que toutes les K classes $c_1, \dots, c_K$ sont équiprobables, auquel cas on a:

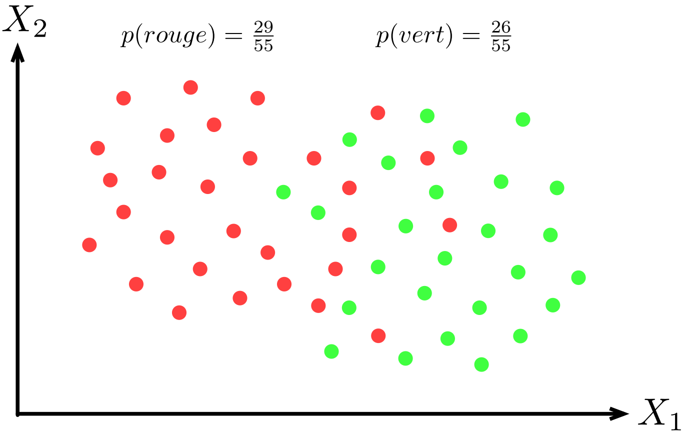
$$p(c_k) = \frac{1}{K} \text{ , pour tout } k = 1, \dots, K.$$

- Ou alors on estime $p(c_k)$ comme la proportion d'éléments du train set S qui sont de la classe c_k, i.e.

$$p(c_k) = \frac{|S_k|}{N} \text{ , pour tout } k = 1, \dots, K.$$

où S_k est le sous-dataset de S formé des éléments de classe c_k.

Estimation des priors



Gaussian Naive Bayes Classifier

- **Estimation des "feature distributions":** l'estimation des $p(X_j = x_j \mid Y = c_k) = p(x_j \mid c_k)$ pour $j = 1, \ldots, p$ $k = 1, \ldots, K$ à partir des data diffère selon la nature des data.
- Si les features X_i sont *continues*, on suppose généralement que chaque $p(X_j \mid Y = c_k)$ suit une *loi normale*, i.e.

$$p(X_j = x_j \mid Y = c_k) \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$$

- On obtient alors un **Gaussian naive Bayes classifier**.

Gaussian Naive Bayes Classifier

- **Estimation des "feature distributions":** l'estimation des $p(X_j = x_j \mid Y = c_k) = p(x_j \mid c_k)$ pour $j = 1, \dots, p$ $k = 1, \dots, K$ à partir des data diffère selon la nature des data.
- Si les features X_i sont *continues*, on suppose généralement que chaque $p(X_j \mid Y = c_k)$ suit une *loi normale*, i.e.

$$p(X_j = x_j \mid Y = c_k) \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$$

- On obtient alors un **Gaussian naive Bayes classifier**.

Gaussian Naive Bayes Classifier

- **Estimation des “feature distributions”:** l'estimation des $p(X_j = x_j \mid Y = c_k) = p(x_j \mid c_k)$ pour $j = 1, \ldots, p$ $k = 1, \ldots, K$ à partir des data diffère selon la nature des data.
- Si les features X_i sont *continues*, on suppose généralement que chaque $p(X_j \mid Y = c_k)$ suit une *loi normale*, i.e.

$$p(X_j = x_j \mid Y = c_k) \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$$

- On obtient alors un **Gaussian naive Bayes classifier**.

Gaussian Naive Bayes Classifier

- Pour tout $j = 1, \ldots, p$ et pour tout $k = 1, \ldots, K$, on estime la moyenne μ_{jk} et la variance σ_{jk}^2 par

$$\mu_{jk} = \frac{1}{|S_k|} \sum_{(\boldsymbol{x},y) \in S_k} x_j \quad \text{et} \quad \sigma_{jk}^2 = \frac{1}{(|S_k| - 1)} \sum_{(\boldsymbol{x},y) \in S_k} (x_j - \mu_{jk})^2$$

  où S_k est le sous-dataset de S formé des éléments qui sont de classe c_k.
- On a donc $2PK$ paramètres (c'est peu !).

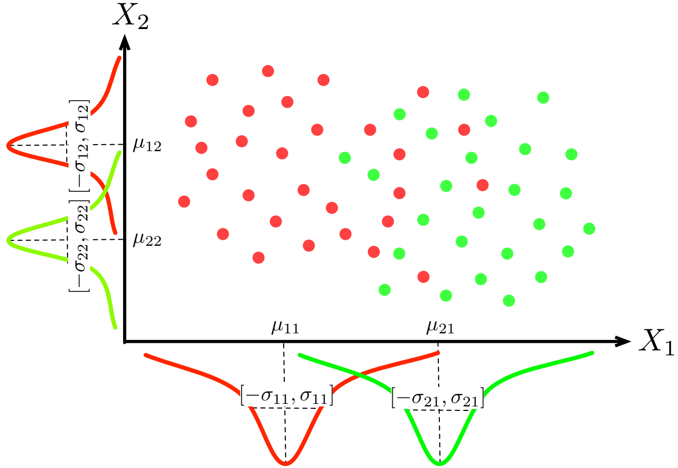
Gaussian Naive Bayes Classifier

- Pour tout $j = 1, \ldots, p$ et pour tout $k = 1, \ldots, K$, on estime la moyenne μ_{jk} et la variance σ_{jk}^2 par

$$\mu_{jk} = \frac{1}{|S_k|} \sum_{(\boldsymbol{x},y) \in S_k} x_j \quad \text{et} \quad \sigma_{jk}^2 = \frac{1}{(|S_k| - 1)} \sum_{(\boldsymbol{x},y) \in S_k} (x_j - \mu_{jk})^2$$

  où S_k est le sous-dataset de S formé des éléments qui sont de classe c_k.
- On a donc $2PK$ paramètres (c'est peu !).

Gaussian Naive Bayes Classifier



Gaussian Naive Bayes Classifier

- Ensuite, puisque $p(X_j = x_j \mid c_k) \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$, on a

$$p(x_j \mid c_k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} e^{-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}}$$

- Ainsi, pour tout point $\boldsymbol{x} = (x_1, \ldots, x_P)$ et toute classe c_k, nous avons tout ce qu'il faut pour calculer les formules du naive Bayes classifier:

$$\begin{aligned} p(c_k \mid \boldsymbol{x}) &\propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \\ \hat{c} &= \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \end{aligned}$$

Gaussian Naive Bayes Classifier

- Ensuite, puisque $p(X_j = x_j \mid c_k) \sim \mathcal{N}(\mu_{jk}, \sigma^2_{jk})$, on a

$$p(x_j \mid c_k) = \frac{1}{\sqrt{2\pi\sigma^2_{jk}}} e^{-\frac{(x_j - \mu_{jk})^2}{2\sigma^2_{jk}}}$$

- Ainsi, pour tout point $\boldsymbol{x} = (x_1, \ldots, x_P)$ et toute classe c_k, nous avons tout ce qu'il faut pour calculer les formules du naive Bayes classifier:

$$\begin{aligned} p(c_k \mid \boldsymbol{x}) &\propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \\ \hat{c} &= \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \end{aligned}$$

Multinomial Naive Bayes Classifier

- Si les features X_i sont *discrètes*, on suppose généralement que le vecteur de features $p(\boldsymbol{X} \mid Y = c_k)$ (et non chaque feature individuelle) suit une *loi multinomiale*, i.e.

$$p(\boldsymbol{X} = \boldsymbol{x} \mid Y = c_k) \sim Multinomial(\pi_{1k}, \ldots, \pi_{Pk})$$

  où π_{jk} représente la probabilité que le j-ième évènement apparaisse dans la classe c_k.

- On obtient alors un **multinomial naive Bayes classifier**.

Multinomial Naive Bayes Classifier

- Si les features X_i sont *discrètes*, on suppose généralement que le vecteur de features $p(\boldsymbol{X} \mid Y = c_k)$ (et non chaque feature individuelle) suit une *loi multinomiale*, i.e.

$$p(\boldsymbol{X} = \boldsymbol{x} \mid Y = c_k) \sim Multinomial(\pi_{1k}, \ldots, \pi_{Pk})$$

  où π_{jk} représente la probabilité que le j-ième évènement apparaisse dans la classe c_k.

- On obtient alors un **multinomial naive Bayes classifier**.

Multinomial Naive Bayes Classifier

- Pour tout $j = 1, \ldots, p$ et pour tout $k = 1, \ldots, K$, on estime la probabilité π_{jk} par

$$\pi_{jk} = \frac{\sum_{\boldsymbol{x} \in S_k} x_j}{\sum_{j'=1}^{P} \sum_{\boldsymbol{x} \in S_k} x_{j'}}$$

  où S_k est le sous-dataset de S formé des éléments qui sont de classe c_k.

- On a donc PK paramètres (c'est peu !).

Multinomial Naive Bayes Classifier

- Pour tout $j = 1, \ldots, p$ et pour tout $k = 1, \ldots, K$, on estime la probabilité π_{jk} par

$$\pi_{jk} = \frac{\sum_{\boldsymbol{x} \in S_k} x_j}{\sum_{j'=1}^{P} \sum_{\boldsymbol{x} \in S_k} x_{j'}}$$

  où S_k est le sous-dataset de S formé des éléments qui sont de classe c_k.

- On a donc PK paramètres (c'est peu !).

Multinomial Naive Bayes Classifier

- Ensuite, puisque $p(\boldsymbol{X} = \boldsymbol{x} \mid c_k) \sim Multinomial(\pi_{1k}, \dots, \pi_{Pk})$, on a

$$p(\boldsymbol{x} \mid c_k) = \left(\frac{\sum_{j=1}^{P} x_j!}{x_1! \cdots x_P!} \right) {\pi_{1k}}^{x_1} \dots {\pi_{Pk}}^{x_P}$$

- Ainsi, pour tout point $\boldsymbol{x} = (x_1, \dots, x_P)$ et toute classe c_k, nous avons tout ce qu'il faut pour calculer les formules du naive Bayes classifier:

$$p(c_k \mid \boldsymbol{x}) \propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

$$\hat{c} = \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k)$$

Multinomial Naive Bayes Classifier

- Ensuite, puisque $p(\boldsymbol{X} = \boldsymbol{x} \mid c_k) \sim Multinomial(\pi_{1k}, \ldots, \pi_{Pk})$, on a

$$p(\boldsymbol{x} \mid c_k) = \left(\frac{\sum_{j=1}^{P} x_j!}{x_1! \cdots x_P!} \right) {\pi_{1k}}^{x_1} \ldots {\pi_{Pk}}^{x_P}$$

- Ainsi, pour tout point $\boldsymbol{x} = (x_1, \ldots, x_P)$ et toute classe c_k, nous avons tout ce qu'il faut pour calculer les formules du naive Bayes classifier:

$$\begin{aligned} p(c_k \mid \boldsymbol{x}) &\propto p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \\ \hat{c} &= \arg\max_{c_k \in C} p(c_k) \prod_{j=1}^{P} p(x_j \mid c_k) \end{aligned}$$

Bibliographie

3Blue1Brown.
Bayes theorem, the geometry of changing beliefs.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R, volume 103 of *Springer Texts in Statistics*.
Springer, New York.

Wikipedia contributors (2023).
Naive bayes classifier — Wikipedia, the free encyclopedia.