Introduction à l'apprentissage automatique

François Denis, Hachem Kadri, Cécile Capponi

Laboratoire d'Informatique Fondamentale de Marseille
Université d'Aix-Marseille

December 14, 2016

Exemples de problèmes traités

Classification supervisée

- *Email/Spam* : écarter automatiquement les annonces publicitaires et autres messages non sollicités.
  Données : des messages réguliers et des SPAMs.
  Classification binaire
- *Reconnaissance des visages, des chiffres, . . .* : reconnaître une personne à partir d'une photo, un code postal à partir d'un manuscrit.
  Données : des images de chiffres ou de visages identifiés.
  Classification n-aire
- *Mots-clés* : attribuer un ou plusieurs mots-clés à des articles de journaux.
  Données : des articles provenant de rubriques diverses.
  Classification multi-étiquette

Objectif : classer correctement de **nouvelles** instances.

Exemples de problèmes traités (suite)

Régression supervisée

- Prédire la température, la pression atmosphérique ou la vitesse du vent en fonction de divers paramètres numériques ou symboliques.
- Dans le problème de détection des SPAMs, associer à un nouveau document la *probabilité* que ce soit un SPAM.
- Prédire le coût d'un client d'une compagnie d'assurances.
- Prédire la durée de vie d'un patient atteint d'une certaine maladie.

Une valeur numérique est attachée à un objet, une situation, un événement. On cherche à prédire cette valeur pour de nouvelles instances.

Exemples de problèmes traités (suite)

Estimation de densité

- Afin de différencier deux auteurs à partir des documents qu'ils ont produits, étudier les fréquences de mots apparaissant dans leurs oeuvres pour estimer la probabilité qu'un nouveau document ait été écrit par l'un ou l'autre auteur.
- La distribution des nucléotides est-elle la même dans les parties codantes et les parties non codantes d'un gène ?
- Une famille de protéines données peut-elle être caractérisée par la distribution de leurs acides aminés ?

On observe un certain nombre d'objets, de situations, d'événements. On cherche à prédire la probabilité d'apparition d'instances semblables.

Exemples de problèmes traités (suite)

Classification non-supervisée (ou clustering ou segmentation)

- Identifier des *profils* parmi les clients d'une entreprise, les usagers des transports en commun ou les spectateurs d'une chaîne de télévision : la fameuse ménagère de plus de 40 ans, les bobos, les couples en voie d'acheter un bien immobilier, les "grands voyageurs", ...
- Une puce à ADN décrit les niveaux d'expression de quelques milliers de gènes prélevés dans divers tissus sains ou cancéreux. Peut-on en déduire des *patterns d'expression*, c'est-à-dire des régularités, des comportements corrélés, ... ?

On observe un certain nombre d'objets, de situations, d'événements. On cherche à en déduire des classes dans lesquelles de nouvelles instances comparables pourront s'insérer.

Quelques ouvrages généraux et autres ressources sur l'apprentissage automatique

- *Apprentissage artificiel*, d'Antoine Cornuejols et Laurent Miclet. En français !
- *The elements of statistical Learning* de Hastie, Tibshirani et Friedman. Plus difficile.
- *Pattern recognition and Machine Learning*, de Christopher M. Bishop, Springer. Plus difficile.
- *All of Statistics, a Concise Course in Statistical Inference*, Larry Wasserman, Springer. Des connaissances statistiques de base pour l'apprentissage automatique.
- Les cours de Philippe Besse à l'INSA de Toulouse.
- De nombreux MOOC, dont ceux de l'université de Stanford : `https://www.coursera.org/learn/machine-learning`.
- La plate-forme logicielle `ScikitLearn` : `http://scikit-learn.org/`

Modélisation de la classification supervisée

- Un ensemble
$$\mathcal{X} = X_1 \times \ldots \times X_n$$
où chaque X_i est le domaine d'un attribut A_i symbolique, numérique ou structuré.
- Un ensemble fini de classes $\mathcal{Y}$
- Une variable aléatoire $Z = (X, Y)$ à valeurs dans $\mathcal{X} \times \mathcal{Y}$
- Soit P_Z la loi de probabilité de Z :
$$P_Z(Z = (x, y)) = P(X = x)P(Y = y|X = x) \text{ si } P(X = x) \neq 0.$$
- Les exemples sont des couples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ tirés selon P_Z.
- Un échantillon S est un multi-ensemble fini d'exemples $\{(x_1, y_1), \ldots, (x_l, y_l)\}$ i.i.d. selon P_Z.

Modélisation de la classification supervisée (suite)

Classifieur :
$$f : \mathcal{X} \to \mathcal{Y}.$$

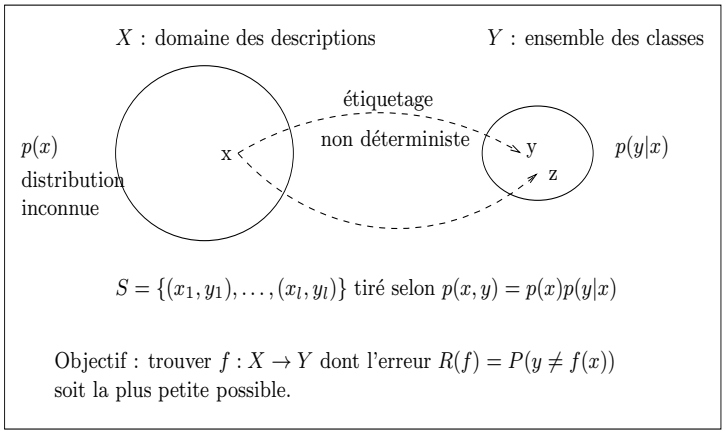
Fonction de perte (loss function)
$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$
$$L(y, f(x)) = \begin{cases} 0 \text{ si } y = f(x) \\ 1 \text{ sinon.} \end{cases}$$

La fonction risque (ou erreur) : espérance mathématique de la fonction de perte.
$$R(f) = \int L(y, f(x)) dP_Z(x, y) = \int_{y \neq f(x)} dP_Z(x, y) = P_Z(y \neq f(x)).$$

Le problème général de la classification supervisée :

étant donné un échantillon $S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, trouver un classifieur f qui minimise le risque $R(f)$.

Modélisation de la classification supervisée (suite)



$S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ tiré selon $p(x, y) = p(x)p(y|x)$

Objectif : trouver $f : X \to Y$ dont l'erreur $R(f) = P(y \neq f(x))$ soit la plus petite possible.

Remarques

1. Le modèle est non déterministe : deux étiquettes différentes peuvent être associées à la même description. En effet,
   - le problème peut être bruité ;
   - l'espace de descriptions peut ne décrire qu'incomplètement une situation complexe.
2. Le problème est non déterministe mais on en cherche une solution déterministe.
3. Le modèle est non paramétrique : aucun modèle spécifique de génération de données n'est présupposé ; aucune contrainte sur l'ensemble des fonctions que l'on doit considérer ni sur le type de dépendances entre fonctions et paramètres.
4. D'autres fonctions de pertes peuvent être considérées. En particulier, on peut envisager des coûts différents selon les erreurs commises.

Quelques règles de classification

- La règle majoritaire : pour toute nouvelle instance, retourner la classe y_{maj} majoritaire, c'est-à-dire pour laquelle $P(y)$ est maximale : pour tout $x \in X$,

$$f_{maj}(x) = \underset{y}{ArgMax} P(y) = y_{maj} \text{ et } R(f_{maj}) = 1 - P(y_{maj}).$$

- La règle du maximum de vraisemblance (*maximum likelihood*) : retourner pour chaque instance x la classe y pour laquelle x est la valeur la plus observée.

$$f_{mv}(x) = \underset{y}{ArgMax} P(x|y).$$

- La règle de Bayes : retourner pour chaque instance x, la classe y dont l'observation est la plus probable, ayant observé x.

$$f_B(x) = \underset{y}{ArgMax} P(y|x).$$

Exemple

Les pièces de monnaie peuvent présenter certaines irrégularités. Une étude montre que

- 33% d'entre elles sont équilibrées,
- 32% sont biaisées en faveur de Pile, avec en moyenne une probabilité de 51% de tomber sur Pile,
- 35% sont biaisées en faveur de Face, avec en moyenne une probabilité de 51% de tomber sur Face.

On jette une pièce, elle tombe sur Pile : que peut-on en déduire ?

Optimalité de la règle de Bayes

Théorème : La règle de décision de Bayes est la règle de risque minimal.

Démonstration Soit f une règle de décision. On a

$$R(f) = \sum_{f(x)\neq y} P(x,y) = \sum_{x\in X} P(x) \sum_{y\neq f(x)} P(y|x) = \sum_{x\in X} P(x)(1 - P(f(x)|x)).$$

On en déduit en particulier que

$$R(f_B) = \sum_{x\in X} P(x)(1 - P(f_B(x)|x)).$$

On sait que $P(f_B(x)|x) \geq P(y|x)$ pour tout $y \in Y$. En particulier, $P(f_B(x)|x) \geq P(f(x)|x)$. On en déduit que

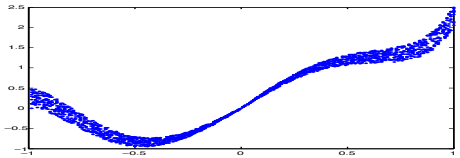
$$R(f) = \sum_{x\in X} P(x)(1 - P(f(x)|x)) \geq \sum_{x\in X} P(x)(1 - P(f_B(x)|x)) = R(f_B).$$

Corollaire :

- si le classifieur de Bayes pouvait être calculé, le problème de la classification supervisée serait résolu !
- le risque du classifieur de Bayes mesure le non-déterminisme du problème à résoudre.

Régression : modélisation

La variable y prend des valeurs continues.



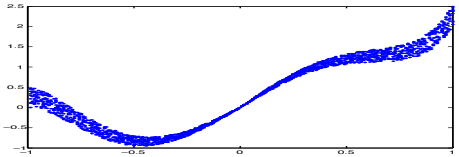
Fonction de perte : l'*écart quadratique* défini par

$$L(y, f(x)) = (y - f(x))^2.$$

Le risque ou l'erreur d'une fonction f : l'*écart quadratique moyen* défini par :

$$R(f) = \int_{X\times Y} (y - f(x))^2 dP(x,y).$$

Régression : modélisation (suite)



Théorème : La fonction $\bar{f}$, moyenne des valeurs observables en x, définie par

$$\bar{f}(x) = \int_Y y dP(y|x)$$

est la fonction de régression de risque minimal.

Estimation de densité : modélisation

On suppose l'existence d'une distribution de probabilité P sur $\mathcal{X}$ (cas discret).
On définit la fonction de perte :

$$L(x, y) = -\log y$$

et pour toute distribution de probabilités P' définie sur $\mathcal{X}$ (cas discret), la fonction de risque :

$$R(P') = \sum_{x\in\mathcal{X}} -\log P'(x) \cdot P(x).$$

Estimation de densité : modélisation (suite)

$$R(P') = \sum_{x \in \mathcal{X}} -\log P'(x) \cdot P(x).$$

Théorème : $R(P')$ est minimal pour $P' = P$.

Exemple :

- $\mathcal{X} = \{Pile, Face\}$, $P(Pile) = 1/3, P(Face) = 2/3$.
- Soit $P'(Pile) = p$ et $P'(Face) = 1 - p$.

$$R(P') = -\log P'(Pile) \cdot P(Pile) - \log P'(Face) \cdot P(Face) = -\frac{1}{3}(\log p + 2\log(1-p)).$$

$$\frac{dR(P')}{dp} = -\frac{1}{3}(\frac{1}{p} - \frac{2}{(1-p)}) = -\frac{1}{3} \cdot \frac{1-3p}{p(1-p)}.$$

$\frac{dR(P')}{dp} = 0$ ssi $p = 1/3$.

L'apprentissage en pratique

On dispose d'un échantillon fini S qu'on suppose i.i.d.

On recherche une fonction f de *classification*, de *régression* ou de *densité* dont le risque $R(f)$ soit le plus faible possible.

Il existe toujours une meilleure solution f_{min} ... inaccessible !

Comment utiliser l'échantillon S pour apprendre une fonction f de risque faible ?

Une idée naturelle : utiliser les données pour estimer $R(f)$.

La notion de risque empirique

Le *risque empirique* $R_{emp}(f)$ d'une fonction f sur l'échantillon $S = \{(x_1, y_1), \ldots (x_l, y_l)\}$ est la moyenne de la fonction de perte calculée sur S :

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i)).$$

$R_{emp}(f)$ est une estimation du risque réel $R(f)$ de f.

La notion de risque empirique (suite)

en classification : $R_{emp}(f)$ est la moyenne du nombre d'erreurs de prédiction de f sur les éléments de S :

$$R_{emp}(f) = \frac{Card\{i | f(x_i) \neq y_i\}}{l}.$$

en régression : $R_{emp}(f)$ est la moyenne des carrés des écarts à la moyenne de f sur S :

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2.$$

en estimation de densité : $R_{emp}(P')$ est l'opposé de la log-vraisemblance de S :

$$R_{emp}(P') = \frac{1}{l}\sum_{i=1}^{l} -\log P'(x_i) = \frac{-1}{l}\log\prod_{i=1}^{l} P'(x_i) = \frac{-1}{l}\log P'(S).$$

Le principe de minimisation du risque empirique

Le *principe inductif de minimisation du risque empirique (ERM)* recommande de

trouver une fonction f qui minimise $R_{emp}(f)$.

- en classification, cela revient à minimiser le nombre d'erreurs commises par f sur l'échantillon ;
- en régression, on retrouve la *méthode des moindres carrés* ;
- en estimation de densité, on retrouve la *méthode du maximum de vraisemblance*.

Problème : la fonction qui minimise $R_{emp}(f)$

- est très simple à trouver,
- n'a aucune valeur de généralisation !

Le principe de minimisation du risque empirique (suite)

- $R_{emp}(f)$ est une estimation du risque réel $R(f)$ de f
- Pour trouver une fonction qui minimise $R(f)$, choisir la fonction f_{emp} qui minimise $R_{emp}(f)$
- la fonction f_{emp} n'a aucune valeur de généralisation !

D'où vient le problème ?

Comme le même ensemble sert à calculer f_{emp} et à estimer son risque, $R_{emp}(f_{emp})$ n'est pas une bonne estimation de $R(f_{emp})$!

Solution : restreindre l'ensemble $\mathcal{F}$ dans lequel la fonction f est recherchée.

- $R_{emp}(f_{emp})$ reste une estimation *biaisée* de $R(f_{emp})$ (trop optimiste)
- mais f_{emp} peut avoir de très bonnes qualités de généralisation si l'ensemble $\mathcal{F}$ est bien choisi.

L'apprentissage en pratique (suite)

Dans la pratique, on cherche une solution dans des ensembles de fonctions $\mathcal{F}$ particuliers :

k-plus proches voisins, arbres de décision, *réseaux de neurones*, *fonctions linéaires*, *fonction polynomiales*, *modèles de Markov cachés*, etc

dont la capacité d'expression doit être réglée en fonction de l'échantillon disponible.

Soit

- f_{min} une fonction qui minimise $R(f)$,
- $f_{opt} \in \mathcal{F}$ une fonction de $\mathcal{F}$ qui minimise $R(f)$,
- $f_{emp} \in \mathcal{F}$ une fonction de $\mathcal{F}$ qui minimise $R_{emp}(f)$,

$$R(f_{emp}) = R(f_{min}) + [R(f_{opt}) - R(f_{min})] + [R(f_{emp}) - R(f_{opt})]$$

Le principe de minimisation du risque empirique (suite)

$$R(f_{emp}) = R(f_{min}) + [R(f_{opt}) - R(f_{min})] + [R(f_{emp}) - R(f_{opt})]$$

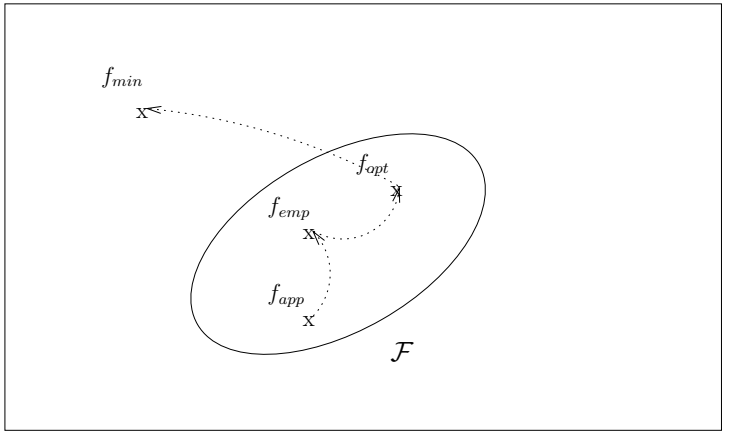
- $R(f_{min})$: incompressible, donne une mesure de la difficulté intrinsèque du problème, du volume de bruit qu'il comporte.
- $R(f_{opt}) - R(f_{min})$: mesure l'adéquation de $\mathcal{F}$ au problème considéré. Si la capacité de $\mathcal{F}$ est réduite, ce terme risque d'être important.
- $R(f_{emp}) - R(f_{opt})$: représente l'erreur liée au principe de minimisation du risque empirique. Si la capacité de $\mathcal{F}$ est trop grande, ce terme risque d'être important.

Niveaux de difficultés en apprentissage

Il y a donc au moins quatre raisons pour lesquelles une méthode d'apprentissage appliquée à un problème particulier peut ne pas donner de résultats satisfaisants :

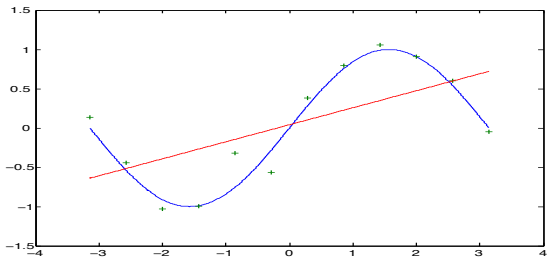
- la *nature non déterministe du problème*,
- la trop *faible expressivité* de l'espace fonctionnel $\mathcal{F}$ choisi,
- la *non consistance du principe ERM* ou plus généralement, du principe choisi pour approcher une fonction optimale dans $\mathcal{F}$,
- la *difficulté à minimiser le risque empirique* (ou plus généralement, à mettre en application le principe choisi).

Niveaux de difficultés en apprentissage (suite)



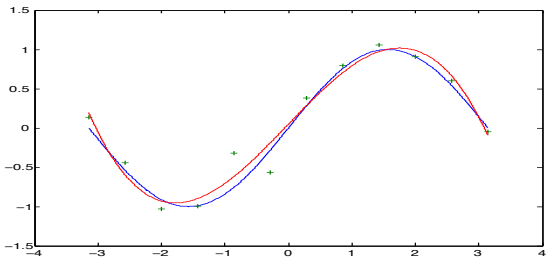
Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 1 qui minimise le risque empirique quadratique.



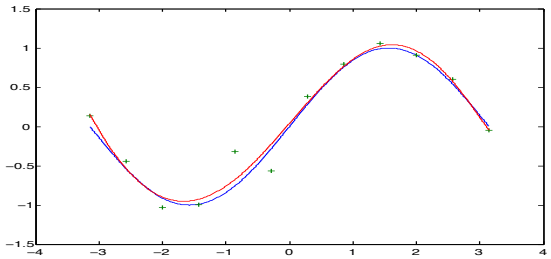
Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 3 qui minimise le risque empirique quadratique.



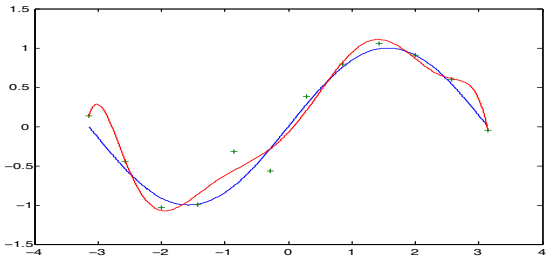
Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 5 qui minimise le risque empirique quadratique.



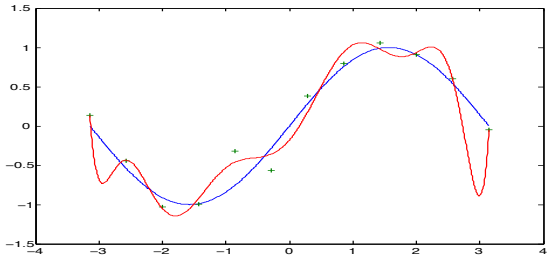
Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 7 qui minimise le risque empirique quadratique.



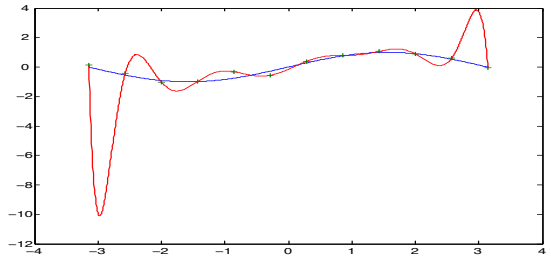
Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 9 qui minimise le risque empirique quadratique.



Estimation de la régression à l'aide de fonctions polynômes.

Données : 11 points sur la courbe $x \mapsto sin(x)$ avec un bruit additif normal d'écart-type 0.2.
En bleu : la courbe $x \mapsto sin(x)$
En rouge : le polynôme de degré 11 qui minimise le risque empirique quadratique.



Annexe : rappels de probabilités discrètes

Soit Ω un ensemble fini ou dénombrable appelé *univers*. Une *probabilité* sur Ω est une application $p : \Omega \to [0,1]$ telle que $\sum_{\omega\in\Omega} p(\omega) = 1$. On appelle *événement* toute partie de Ω ; pour tout événement A, on définit $p(A) = \sum_{\omega\in A} p(\omega)$.

- si les $(A_i)_{i\in I}$ sont incompatibles 2 à 2,

$$p\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} p(A_i)$$

- $p(A\cup B) = p(A) + p(B) - p(A\cap B)$
- $p(\overline{A}) = 1 - p(A)$ où $\overline{A} = \Omega \setminus A$,
- $A \subseteq B \Rightarrow p(A) \leq p(B)$.

Rappels de probabilités discrètes (suite)

Soit A un événement tel que $p(A) \neq 0$, alors

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

est une probabilité sur Ω.
Formule de Bayes :

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

Si les $(A_i)_{i\in I}$ sont incompatibles 2 à 2 et vérifient $\cup_{i\in I} A_i = \Omega$, alors pour tout B, $p(B) = \sum_{i\in I} p(B|A_i)p(A_i)$.

Rappels de probabilités discrètes (suite)

Les événements A et B sont *indépendants* ssi ils vérifient l'une des conditions suivantes

1. $p(A|B) = p(A)$
2. $p(B|A) = p(B)$
3. $p(A\cap B) = p(A)p(B)$.

Exercices

1. On tire 10 fois de suite une pièce équilibrée.
   1. Quelle est la probabilité qu'elle tombe 10 fois sur pile ?
   2. Quelle est la probabilité qu'elle tombe la dixième fois sur Pile sachant qu'elle est déjà tombée les 9 premières fois sur Pile.
2. On jette 10 fois un dé équilibré. Quelle est la probabilité qu'il tombe au moins une fois sur 6.
3. On jette un dé.
   1. Les événements : *le nombre tiré est pair* et *le nombre tiré est* ≤ 3 sont ils indépendants ?
   2. Quelle est la probabilité que les deux nombres soient différents sachant que le maximum des deux nombres tirés est égal à 5.