Rappel : probabilité, espérance et conditionnement

+ formule de Bayes, biais–variance et dimension de Vapnik–Chervonenkis

Oulad Ali Ayoub · TD X-HEC · APM_51434_EP · École Polytechnique

Dans toute la fiche, les variables aléatoires sont **discrètes** : X prend ses valeurs dans un ensemble fini ou dénombrable $\mathcal{X} \subset \mathbb{R}$ et $p_X(x) = \mathbb{P}(X = x)$.

1 Lire un énoncé : qu'est-ce qui est aléatoire, qu'est-ce qui est fixé ?

Avant de calculer, on repère ce qui est aléatoire. C'est ce qui décide si une quantité peut sortir d'une espérance, et sur quoi porte chaque $\mathbb{E}$.

Les indices dans l'énoncé

Aléatoire si l'énoncé dit ou écrit…	**Fixé** si l'énoncé dit ou écrit…
« *random variable* », « let X be a r.v. », « $X \sim P$ » (le symbole $\sim$ se lit « suit la loi »)	« *fixed* », « *constant* », « *deterministic* », « *given* »
« *drawn* », « *sampled* », « i.i.d. », « *random dataset* D », « *noise* ε »	« let $x \in \mathcal{X}$ », « for a fixed x », « for every x », « let c be a constant »
« possibly dependent on X » : C est aléatoire, mais connu dès qu'on connaît X	les valeurs numériques de l'énoncé : $\mathbb{P}(D = 1) = 0{,}01$, sensibilité 0,99
tout ce qui est calculé à partir d'un objet aléatoire : $g(X)$, $X + Y$, $\hat{f}_D(x)$, $\mathbb{E}[X \mid Y]$	les *fonctions* $f^\star, g, h, \varphi$; et un jeu de données écrit explicitement, comme dans l'exercice 3.7

Ce que font $\mathbb{E}$, Var, $\mathbb{P}$

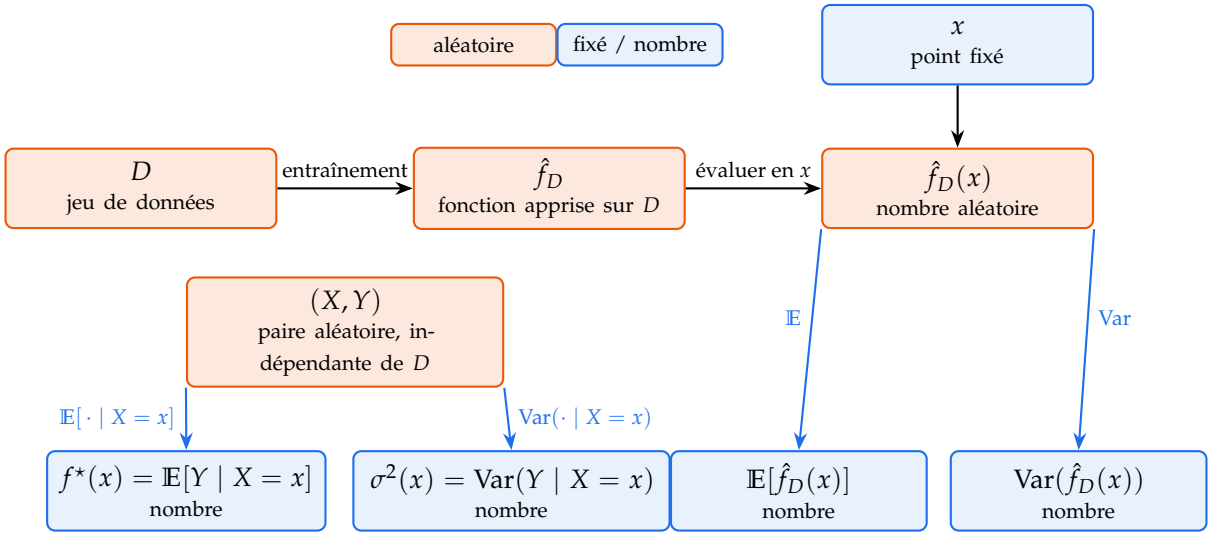
— $\mathbb{E}[\,\cdot\,]$, $\mathrm{Var}(\,\cdot\,)$, $\mathbb{P}(\,\cdot\,)$ prennent une quantité aléatoire et rendent un **nombre** (fixé). Donc, pour x fixé, $\mathbb{E}[\hat{f}_D(x)]$ et $\mathrm{Var}(\hat{f}_D(x))$ sont des nombres, et $\mathbb{E}\big[\mathbb{E}[X]\big] = \mathbb{E}[X]$.

— $\mathbb{E}[X \mid Y = y]$ est un **nombre** ; $\mathbb{E}[X \mid Y]$ est une **variable aléatoire** (une fonction de Y). Même chose pour $\mathrm{Var}(X \mid Y = y)$ et $\mathrm{Var}(X \mid Y)$.

— Sans indice, $\mathbb{E}$ porte sur **tout ce qui est aléatoire entre les crochets**. Dans $\mathbb{E}\big[(Y - \hat{f}_D(x))^2 \mid X = x\big]$, il y a deux sources d'aléa : Y (sachant $X = x$) et le jeu de données D ; le point x, lui, est fixé.

La lettre aide, mais ne prouve rien

En général les majuscules ($X, Y, Z, D, T, \varepsilon$) sont aléatoires et les minuscules (x, y, c) sont fixées. Mais $\mathcal{X}, \mathcal{H}$ sont des ensembles, P est une loi, et $\hat{f}_D$ porte une minuscule alors qu'elle est aléatoire (elle dépend de D). C'est ce que dit l'énoncé qui décide.



Exercice 3.5 : tout ce qui est en orange est aléatoire, tout ce qui est en bleu est un nombre. Les trois termes de la décomposition sont des nombres.

Dans le PC3 : qui est aléatoire, qui est fixé ?

	Aléatoire	**Fixé (nombre, fonction, point)**
3.1	D (malade ou non), T (test)	$\mathbb{P}(D = 1) = 0{,}01$, sensibilité 0,99, spécificité 0,95, et $\mathbb{P}(D = 1 \mid T = 1)$ (un nombre)
3.2, 3.3	X, Y, $g(Y)$, $\mathbb{E}[X \mid Y]$, $\mathrm{Var}(X \mid Y)$	$\mathbb{E}[X \mid Y = y]$, $\mathbb{E}[X]$, $\mathrm{Var}(X)$, la fonction g, $\mathbb{E}[\mathrm{Var}(X \mid Y)]$
3.4	Z ; C (« possibly dependent on X »)	la constante c, $\mathbb{E}[Z]$, $\mathrm{Var}(Z)$, le nombre $\mathbb{E}[(Z - c)^2]$. Sachant X, C est connu : on le traite comme une constante
3.5	(X, Y), D, $\hat{f}_D$, $\hat{f}_D(x)$	x, $f^\star(x)$, $\sigma^2(x)$, $\mathbb{E}[\hat{f}_D(x)]$, $\mathrm{Var}(\hat{f}_D(x))$, $\mathcal{E}(x)$
3.6	rien (pas de probabilités)	le classifieur h, l'arbre, les points de $\{0, 1\}^d$
3.7	rien : l'algorithme ID3 est déterministe	le jeu de données, donné explicitement

1.1 Mini-glossaire des mots de l'énoncé

Mot	**Sens**
pmf	*probability mass function* : $p_X(x) = \mathbb{P}(X = x)$, avec $\sum_x p_X(x) = 1$.
carré intégrable	$\mathbb{E}[Z^2] < \infty$. C'est ce qui garantit que $\mathrm{Var}(Z)$ existe et est finie.
i.i.d.	indépendantes et identiquement distribuées : les points du jeu de données sont tirés indépendamment, suivant la même loi.
jeu de données D	$D = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$, en général i.i.d. Il est *aléatoire* : un autre tirage donnerait un autre D.
estimateur $\hat{f}_D$	règle qui à un jeu de données D associe une fonction. Comme D est aléatoire, $\hat{f}_D$ l'est aussi ; le chapeau signale qu'on l'estime à partir des données.
$f^\star(x) = \mathbb{E}[Y \mid X = x]$	la *vraie* moyenne conditionnelle, la cible de l'estimateur. C'est une fonction fixée, inconnue de l'algorithme.

Mot	Sens
bruit ε	ce qui reste de Y une fois la moyenne retirée : $\varepsilon = Y - f^\star(X)$, avec $\mathbb{E}[\varepsilon \mid X] = 0$.
classe d'hypothèses $\mathcal{H}$	l'ensemble des fonctions que l'algorithme peut produire (par exemple, tous les arbres de profondeur au plus k).
profondeur, hauteur	la profondeur est le nombre de tests successifs le long d'un chemin ; la hauteur compte les niveaux, feuilles comprises, donc d tests donnent une hauteur $d+1$.
souche (*decision stump*)	arbre de profondeur 1 : un seul test, deux feuilles.
ID3 / CART	algorithmes gloutons qui construisent l'arbre un test à la fois, en choisissant à chaque étape le meilleur test local.

2 Espérance et variance

Définitions (sous forme de sommes)

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p_X(x), \qquad \mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\, p_X(x),$$

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \sum_{x \in \mathcal{X}} \big(x - \mathbb{E}[X]\big)^2 p_X(x) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Règles à connaître

- **Linéarité** (toujours vraie, même sans indépendance) : $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$.
- $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.
- Si X et Y sont **indépendantes** : $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ et $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Variance d'une somme

Sans indépendance, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$. Le terme de covariance disparaît seulement si X et Y sont indépendantes (ou non corrélées).

2.1 Comment retenir $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$: l'inégalité de Jensen

Inégalité de Jensen

Si φ est une **fonction convexe**, alors $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.
Avec $\varphi(x) = x^2$, qui est convexe : $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$. Donc $\mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$, ce qui est cohérent avec $\mathrm{Var}(X) \geq 0$. Pour savoir lequel des deux termes est le plus grand, il suffit de se souvenir que le carré est convexe.

Fonction convexe $\neq$ ensemble convexe

Une *fonction* φ est convexe si $\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y)$ pour tout $t \in [0,1]$. Un *ensemble* C est convexe si $x, y \in C \Rightarrow tx + (1-t)y \in C$. Ce ne sont pas les mêmes objets : Jensen parle d'une fonction, alors que le fait qu'une moyenne reste dans l'ensemble considéré (voir §5) parle d'un ensemble.

3 Conditionnement

Définitions (sous forme de sommes)

Pour deux variables aléatoires X, Y :

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \qquad \mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x\, p_{X|Y}(x \mid y),$$

$$\mathrm{Var}(X \mid Y = y) = \sum_{x \in \mathcal{X}} \big(x - \mathbb{E}[X \mid Y = y]\big)^2 p_{X|Y}(x \mid y) = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2.$$

$\mathbb{E}[X \mid Y]$ est la variable aléatoire $\varphi(Y)$, où $\varphi(y) = \mathbb{E}[X \mid Y = y]$: **c'est une fonction de** Y.

Sur un tableau de données : conditionner par $Y = 2$, c'est garder les lignes où $y = 2$

ligne	x	y
1	4	2
2	1	1
3	8	3
4	3	2
5	2	1
6	9	3
7	4	2
8	3	1
9	7	3
10	5	2

— **Espérance** : $\mathbb{E}[X]$ est la moyenne de *toute* la colonne x : $\dfrac{4+1+8+3+2+9+4+3+7+5}{10} = 4{,}6$.

— **Conditionner par** $Y = 2$: on ne garde que les lignes orange (celles où $y = 2$) et on fait la moyenne des x restants : $\mathbb{E}[X \mid Y = 2] = \dfrac{4+3+4+5}{4} = 4$.

— **Sur ces lignes, y est constant** : il vaut toujours 2, donc il n'y a pas de variance sur y.

$$\mathbb{E}[Y \mid Y = 2] = 2, \qquad \mathrm{Var}(Y \mid Y = 2) = 0,$$
$$\mathbb{E}[g(Y) \mid Y = 2] = g(2).$$

Une quantité qui dépend de y sort de la moyenne : $\mathbb{E}[XY \mid Y = 2] = \dfrac{8+6+8+10}{4} = 8 = 2 \cdot \mathbb{E}[X \mid Y = 2]$.

$\mathbb{E}[X \mid Y]$ **est une fonction de** Y : on fait la même chose pour chaque valeur de y.

valeur de y	1	2	3
$\mathbb{E}[X \mid Y = y]$	2	4	8
$\mathbb{P}(Y = y)$	0,3	0,4	0,3

Propriété de la tour : $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = 0{,}3 \times 2 + 0{,}4 \times 4 + 0{,}3 \times 8 = 4{,}6 = \mathbb{E}[X]$.

Quand on conditionne par X, toute fonction de X est une constante

Si g est une fonction, alors $g(X)$ est connu quand on connaît X : on le traite comme une constante.

$$\mathbb{E}[g(X) \mid X] = g(X),$$
$$\mathbb{E}[Y\, g(X) \mid X] = g(X)\, \mathbb{E}[Y \mid X],$$
$$\mathbb{E}[\mathbb{E}[Y \mid X] \mid X] = \mathbb{E}[Y \mid X] \qquad (\text{car } \mathbb{E}[Y \mid X] \text{ est une fonction de } X).$$

Ce qui ne dépend pas de X (ici Y) reste à l'intérieur de l'espérance conditionnelle.

Autres règles

- **Linéarité** : $\mathbb{E}[aY + bZ \mid X] = a\,\mathbb{E}[Y \mid X] + b\,\mathbb{E}[Z \mid X]$.
- **Indépendance** : si X et Y sont indépendantes, $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ (et $\mathrm{Var}(X \mid Y) = \mathrm{Var}(X)$). Conditionner par quelque chose d'indépendant n'apprend rien.
- **Propriété de la tour** : $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$.
- **Orthogonalité** : pour toute fonction g, $\mathbb{E}\big[(X - \mathbb{E}[X \mid Y])\,g(Y)\big] = 0$. Autrement dit $\mathbb{E}[Xg(Y)] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\,g(Y)\big]$.
- **Variance totale** : $\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big)$.

4 Formule de Bayes : prior, vraisemblance, posterior

C'est l'exercice 3.1. On a une **hypothèse** ($D = 1$: la personne est malade) et une **observation** ($T = 1$: le test est positif). Bayes sert à passer de « que vaut l'observation si l'hypothèse est vraie ? » à « que vaut l'hypothèse maintenant que j'ai vu l'observation ? ».

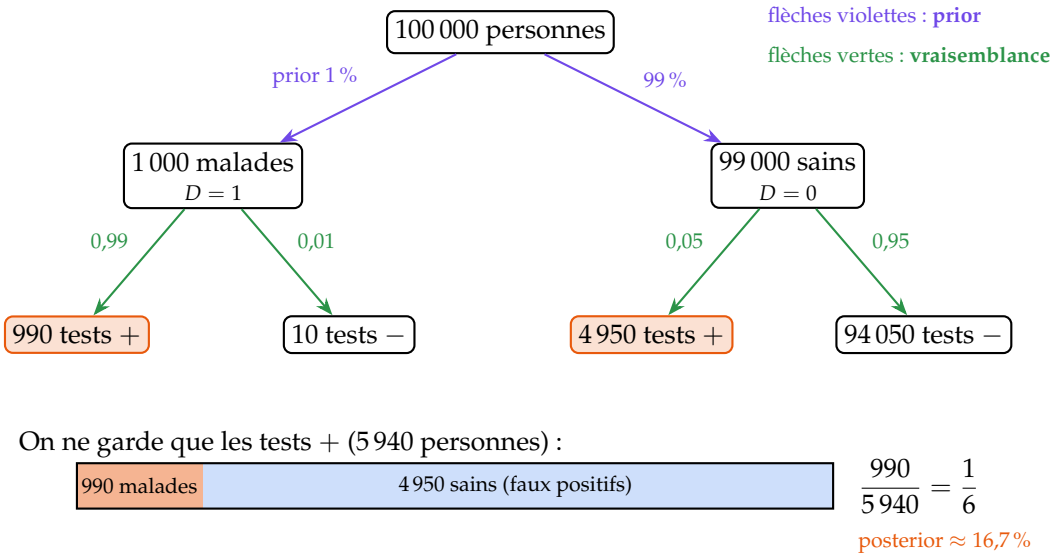
Le vocabulaire

Mot	Notation	Question posée	Exercice 3.1
prior (a priori)	$\mathbb{P}(D = 1)$	Que croit-on *avant* d'avoir vu la donnée ?	1 % (la prévalence)
vraisemblance (*likelihood*)	$\mathbb{P}(T = 1 \mid D = d)$	Si l'hypothèse d est vraie, avec quelle probabilité voit-on la donnée ?	0,99 si $d = 1$, 0,05 si $d = 0$
évidence	$\mathbb{P}(T = 1)$	Avec quelle probabilité voit-on la donnée, toutes hypothèses confondues ?	0,0594
posterior (a posteriori)	$\mathbb{P}(D = 1 \mid T = 1)$	Que croit-on *après* avoir vu la donnée ?	$1/6 \approx 16{,}7\,\%$

$$\underbrace{\mathbb{P}(D = 1 \mid T = 1)}_{\text{posterior}} = \frac{\overbrace{\mathbb{P}(T = 1 \mid D = 1)}^{\text{vraisemblance}}\ \overbrace{\mathbb{P}(D = 1)}^{\text{prior}}}{\underbrace{\mathbb{P}(T = 1)}_{\text{évidence}}},$$

$$\mathbb{P}(T = 1) = \sum_{d \in \{0,1\}} \mathbb{P}(T = 1 \mid D = d)\,\mathbb{P}(D = d).$$

En une phrase : **posterior** $\propto$ **vraisemblance** $\times$ **prior**. L'évidence ne sert qu'à normaliser, pour que $\mathbb{P}(D = 1 \mid T = 1) + \mathbb{P}(D = 0 \mid T = 1) = 1$.



Fréquences naturelles : le posterior est la part de malades parmi les tests positifs. Il est petit parce que les faux positifs viennent d'une population 99 fois plus grande.

Pourquoi le posterior est petit quand le prior est petit (3.1.3) et la mise à jour des cotes (3.1.4)

Le numérateur de Bayes est $\mathbb{P}(T=1 \mid D=1)\mathbb{P}(D=1) = 0{,}99 \times 0{,}01$, petit parce que $\mathbb{P}(D=1)$ est petit. Le dénominateur contient en plus $\mathbb{P}(T=1 \mid D=0)\mathbb{P}(D=0) = 0{,}05 \times 0{,}99$, qui domine : un petit taux de faux positifs appliqué à presque toute la population donne plus de cas qu'un grand taux de vrais positifs appliqué à un petit groupe.
La **cote** (*odds*) d'un événement de probabilité p est $p/(1-p)$. Diviser les deux formes de Bayes ($D=1$ et $D=0$) fait disparaître l'évidence :

$$\underbrace{\frac{\mathbb{P}(D=1 \mid T=1)}{\mathbb{P}(D=0 \mid T=1)}}_{\text{cote a posteriori}} = \underbrace{\frac{\mathbb{P}(T=1 \mid D=1)}{\mathbb{P}(T=1 \mid D=0)}}_{\text{rapport de vraisemblance}} \times \underbrace{\frac{\mathbb{P}(D=1)}{\mathbb{P}(D=0)}}_{\text{cote a priori}} = \frac{0{,}99}{0{,}05} \times \frac{1}{99} = 19{,}8 \times \frac{1}{99} = 0{,}2.$$

Une cote de 0,2 correspond à $p = \frac{0{,}2}{1+0{,}2} = \frac{1}{6}$. Le test multiplie la cote par 19,8, ce qui ne suffit pas à compenser une cote de départ de $1/99$.

Deux erreurs classiques

- **Inverser le conditionnement.** Une sensibilité $\mathbb{P}(T=1 \mid D=1) = 0{,}99$ ne veut pas dire $\mathbb{P}(D=1 \mid T=1) = 0{,}99$. Les deux sont reliés par Bayes, et le prior intervient.
- **Prendre la vraisemblance pour une loi sur l'hypothèse.** À d fixé, $\sum_t \mathbb{P}(T=t \mid D=d) = 1$. Mais en fonction de d, $0{,}99 + 0{,}05 \neq 1$: la vraisemblance n'est pas une probabilité sur d. C'est le produit avec le prior, puis la normalisation, qui donne une loi.

Le même mécanisme dans la suite du PC

Pour une cible binaire $Y \in \{0,1\}$ et des entrées X, $f^\star(x) = \mathbb{E}[Y \mid X=x] = \mathbb{P}(Y=1 \mid X=x)$: la fonction de régression est le **posterior de la classe**, sachant l'entrée. Le prior est $\mathbb{P}(Y=1)$, la proportion de la classe 1 dans la population. Une feuille d'arbre prédit la classe majoritaire de ses points d'entraînement, et la proportion de 1 dans la feuille estime ce posterior. En statistique bayésienne, on fait la même chose avec un paramètre θ à la place de l'hypothèse D : $p(\theta \mid \text{données}) \propto p(\text{données} \mid \theta)\, p(\theta)$.

5 Décomposition biais–variance

5.1 Le cas le plus simple : prédire Z par une constante c

Exercice 3.4

Si Z est de carré intégrable et c est une constante,

$$\mathbb{E}[(Z-c)^2] = \mathrm{Var}(Z) + \big(\mathbb{E}[Z]-c\big)^2.$$

(On ajoute et on retire $\mathbb{E}[Z]$; le terme croisé s'annule car $\mathbb{E}[Z-\mathbb{E}[Z]]=0$.)
Lecture : $\mathbb{E}[Z]$ est la meilleure constante au sens des moindres carrés, et $\mathrm{Var}(Z)$ est l'erreur qu'on ne peut pas réduire. Toute autre constante c paie en plus la pénalité $(\mathbb{E}[Z]-c)^2$.

5.2 La décomposition pour un estimateur $\hat{f}_D$

On observe $Y = f^\star(X)+\varepsilon$ avec $\mathbb{E}[\varepsilon \mid X]=0$ et $\mathrm{Var}(\varepsilon \mid X=x)=\sigma^2(x)$. L'estimateur $\hat{f}_D$ est construit sur un jeu de données D, indépendant du bruit du point (x,Y) sur lequel on évalue. En un point x fixé,

$$\boxed{\mathbb{E}\big[(Y-\hat{f}_D(x))^2\big] = \underbrace{\sigma^2(x)}_{\text{bruit}} + \underbrace{\big(\mathbb{E}[\hat{f}_D(x)]-f^\star(x)\big)^2}_{\text{biais}^2} + \underbrace{\mathrm{Var}\big(\hat{f}_D(x)\big)}_{\text{variance}}}$$

Preuve en deux lignes. Écrire $Y-\hat{f}_D(x) = \varepsilon + (f^\star(x)-\hat{f}_D(x))$. Le terme croisé est nul (bruit centré et indépendant de $\hat{f}_D$), ce qui donne $\sigma^2(x)+\mathbb{E}[(\hat{f}_D(x)-f^\star(x))^2]$. On applique ensuite l'exercice 3.4 avec $Z=\hat{f}_D(x)$ et $c=f^\star(x)$.

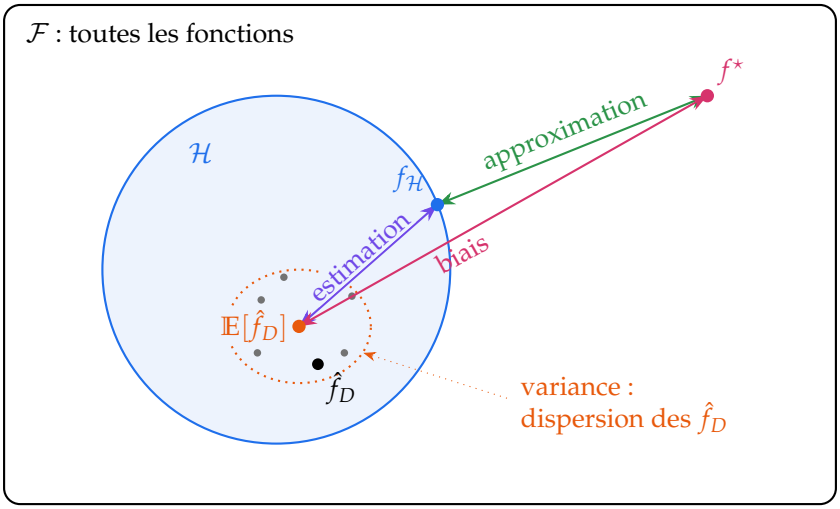
Il n'y a aucune classe d'hypothèses dans cette formule

La décomposition est valable pour *n'importe quel* estimateur. La classe d'hypothèses $\mathcal{H}$ n'intervient que lorsqu'on veut *interpréter* le biais.

5.3 Le rôle de la classe d'hypothèses $\mathcal{H}$

On note $\mathcal{F}$ l'ensemble de toutes les fonctions, $\mathcal{H}\subset\mathcal{F}$ la classe d'hypothèses, et $f_{\mathcal{H}}$ la **projection** de $f^\star$ sur $\mathcal{H}$, c'est-à-dire la meilleure fonction de $\mathcal{H}$ pour approcher $f^\star$ (en norme $L^2(P_X)$). Le biais se coupe en deux morceaux, comme sur le dessin :

$$\underbrace{\mathbb{E}[\hat{f}_D]-f^\star}_{\text{biais}} = \underbrace{(f_{\mathcal{H}}-f^\star)}_{\text{erreur d'approximation}} + \underbrace{\big(\mathbb{E}[\hat{f}_D]-f_{\mathcal{H}}\big)}_{\text{biais d'estimation}}.$$



Par l'inégalité triangulaire, puis par $(a+b)^2 \leq 2a^2 + 2b^2$:

$$\left|\mathbb{E}[\hat{f}_D] - f^\star\right| \leq \underbrace{\left|f_{\mathcal{H}} - f^\star\right|}_{\text{approximation}} + \underbrace{\left|\mathbb{E}[\hat{f}_D] - f_{\mathcal{H}}\right|}_{\text{estimation}}, \qquad \text{biais}^2 \leq 2\,(\text{approximation})^2 + 2\,(\text{estimation})^2.$$

(Le même calcul vaut en norme $L^2(P_X)$ après intégration en x.)

Grossièrement

— **Grande classe** $\mathcal{H}$: l'erreur d'approximation diminue, donc le *majorant* du biais diminue ; la variance augmente.

— **Petite classe** $\mathcal{H}$: la variance diminue ; l'erreur d'approximation, donc le majorant du biais, augmente.

Attention à ce qui diminue vraiment

C'est l'*erreur d'approximation* (donc un majorant du biais) qui diminue avec la taille de la classe, pas forcément le biais lui-même en un point donné. Si $\mathcal{H}$ est un espace vectoriel (régression linéaire) et $f_{\mathcal{H}}$ la projection orthogonale en norme $L^2(P_X)$, Pythagore donne même l'égalité $\|\mathbb{E}\hat{f}_D - f^\star\|^2 = \|f_{\mathcal{H}} - f^\star\|^2 + \|\mathbb{E}\hat{f}_D - f_{\mathcal{H}}\|^2$, sans facteur 2. Sur le dessin, $\mathbb{E}[\hat{f}_D]$ est placé dans le cercle ; c'est garanti si $\mathcal{H}$ est un *ensemble convexe*, mais pas en général (une moyenne d'arbres n'est pas un arbre).

5.4 Schéma : compromis biais–variance selon la profondeur d'un arbre

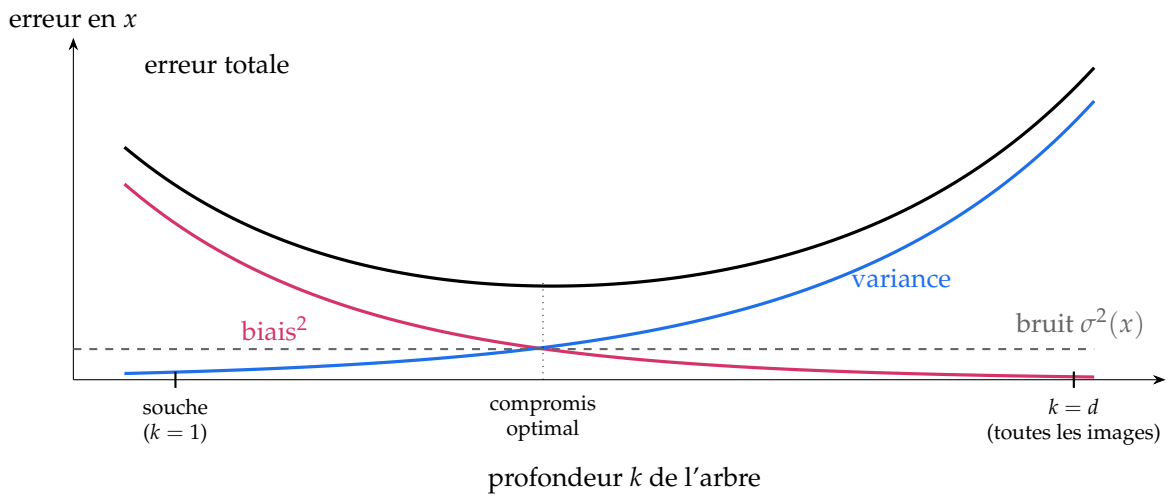


Schéma qualitatif : les courbes ne sont pas des mesures. Le bruit $\sigma^2(x)$ ne dépend pas de l'algorithme.

Souche contre arbre de profondeur d (3.6.3)

— **Souche** (profondeur 1) : un seul test sur un seul pixel, donc très peu de fonctions possibles. Il y en a $2d+2$ (les deux constantes, et pour chaque coordonnée x_i et sa négation), donc VC $\leq \log_2(2d+2)$. La classe est petite : **biais élevé** (elle ne peut pas approcher une fonction compliquée), **variance faible** (changer D change peu la souche choisie).

— **Profondeur** d : toutes les fonctions de $\{0,1\}^d$ dans $\{0,1\}$ sont réalisables, VC $= 2^d$. La classe est énorme : **biais faible** (l'erreur d'approximation est nulle), **variance élevée** (chaque feuille ne contient que quelques points, et l'arbre change beaucoup quand D change).

— Entre les deux, la profondeur règle le compromis : c'est le paramètre qu'on choisit sur un jeu de validation.

6 Dimension de Vapnik–Chervonenkis

Définition

Soit $\mathcal{H}$ un ensemble de classifieurs $h : \mathcal{X} \to \{0,1\}$. Des points $x_1, \ldots, x_m$ sont **pulvérisés** (*shattered*) par $\mathcal{H}$ si, pour tout choix d'étiquettes $(y_1, \ldots, y_m) \in \{0,1\}^m$, il existe $h \in \mathcal{H}$ avec $h(x_i) = y_i$ pour tout i.

$\mathrm{VC}(\mathcal{H}) = \max\{m : \text{ il existe } m \text{ points pulvérisés par } \mathcal{H}\}$ $\quad (= +\infty$ s'il n'y a pas de maximum$)$.

Idée : la dimension VC est la taille du plus grand jeu de données sur lequel on peut **interpoler n'importe quelles étiquettes**. C'est « il existe un jeu de m points », pas « pour tout jeu de m points ».

Classe trop riche

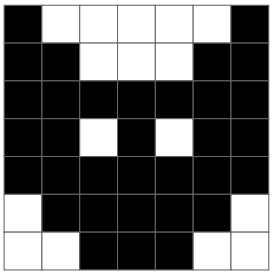
Si $\mathrm{VC}(\mathcal{H})$ est plus grand que le nombre n de données, $\mathcal{H}$ peut s'ajuster à n'importe quelles étiquettes, bruit compris : c'est le risque de surapprentissage.

6.1 Exemple : « chat » ou « pas chat » avec un arbre

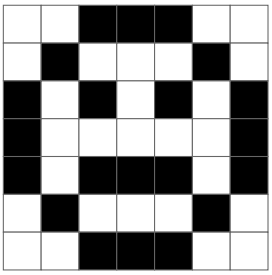
Une image en noir et blanc de 7×7 pixels est un vecteur de $d = 49$ pixels binaires (0 = blanc, 1 = noir) : il y a 2^{49} images possibles. On veut un classifieur qui répond « chat » ou « pas chat ». Les pixels sont numérotés ligne par ligne.

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

numéro des pixels
(lecture ligne par ligne)

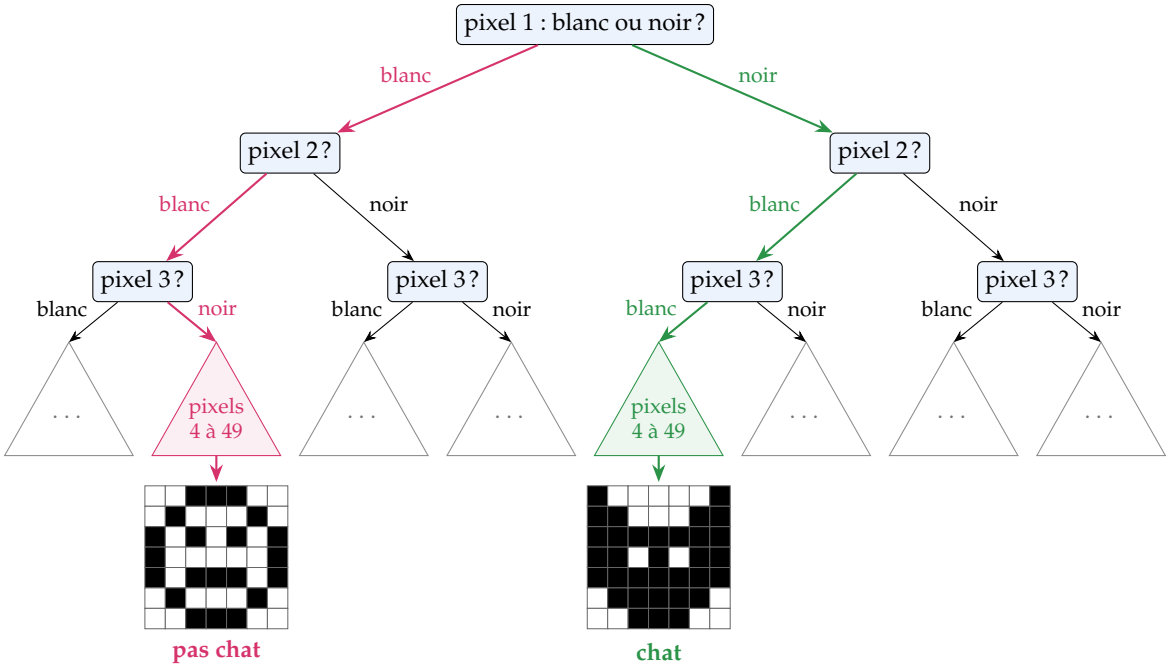


chat
(dessin à la main)



pas chat
(cercle, deux yeux, une bouche)

Construction de l'arbre. À la racine, on regarde le pixel 1 et on demande : est-il blanc ou noir ? Selon la réponse on descend à gauche (blanc) ou à droite (noir). On regarde ensuite le pixel 2, puis le pixel 3, et ainsi de suite jusqu'au pixel 49. À la fin du chemin on arrive dans une feuille qui correspond à *une seule* image : celle dont les 49 pixels sont exactement ceux qu'on a rencontrés.



Dans chaque feuille on écrit l'étiquette qu'on veut. Le dessin de chat (pixel 1 noir, pixel 2 blanc, pixel 3 blanc, etc.) arrive dans une feuille étiquetée « chat ». L'image du cercle avec deux yeux et une bouche n'est pas un chat : sa feuille (pixel 1 blanc, pixel 2 blanc, pixel 3 noir, etc.) reçoit « pas chat ». On fait la même chose pour **toutes** les 2^{49} images possibles : l'arbre contient toutes les images, une par feuille.

Conclusion

Avec d pixels, l'arbre a d tests successifs (profondeur d, soit $d+1$ niveaux en comptant les feuilles) et 2^d feuilles, une par image. Quelles que soient les étiquettes qu'on veut donner à un jeu d'images, il suffit de les écrire dans les feuilles correspondantes : il existe un arbre de profondeur d qui les reproduit. La classe des arbres de profondeur d pulvérise donc les 2^d images, et $\mathrm{VC} = 2^d$.

Un tel arbre apprend le jeu d'entraînement par cœur, bruit compris, mais il ne généralise pas : pour une image qu'on n'a pas vue, l'étiquette de sa feuille a été choisie arbitrairement. C'est le surapprentissage. Il faut donc contrôler la profondeur (ou utiliser une forêt).