Arbre de décision :
Entropie ⇒ $H(X) = -\sum_{i=1}^{n} P(X = x_i).\log(P(X = x_i))$, entropie grande ↔ désordre , nulle ↔ pas d'alé
Entropie cond. ⇒ $H(Y|X) = \sum_i P(X = x_i).H(Y|X = x_i)$
Gain d'information ⇒ $I(T,Y) = H(Y) - H(Y|T)$ à maximiser (donc $H(Y|T)$ à minimiser)

Rappels ⇒ $\mathbb{E}[X] = \int_x x.p(x).dx$, $V[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$, Bayes ⇒ $p(y|x) = p(x|y).p(y)/p(x)$
Conditionnellement ⇒ $p(x|y) = p(x,y)/p(y)$, Gauss-Markov ⇒ pour $X \geq 0, \epsilon > 0$, $P_r(X \geq \epsilon) \leq \mu/\epsilon$
Tchebychev ⇒ $P_r(|X - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2$, Hoeffding ⇒ $X_i \in [a,b], P_r(|\bar{X}_n - \mu| \geq \epsilon) \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$

Classifieur bayésien :
Si on dispose de $p(y)$ et $p(x|y)$, $p(y,x) = p(y|x).p(x) = p(x|y).p(y)$,
$p(x) = p(x|y_+).p(y_+) + p(x|y_-).p(y_-)$ donc $p(y|x) = p(x|y).p(y) / p(x)$.
Décision bayésienne ⇒ $f(x) = \text{argmax}_y\, p(y|x) = \text{argmax}_y\, p(x|y).p(y) / p(x)$
Notion d'erreur/perte ⇒ fct. de perte → $\ell(f(x),y) = 1$ si $f(x) \neq y$, 0 sinon ($f(x) = y$)
Risque associé ⇒ $R(y_i|x) = \sum_j \ell(y_i,y_j).p(y_j|x) = 1 - P(y_i|x)$ (0-1 loss)
Calcul de l'erreur ⇒ $P(\text{erreur}|x) = \min(P(y_+|x), P(y_-|x))$
Classif. bayésien réduit le risque avec f, meilleur classif. possible, $p(x|y)$ très rare ...

Estim. histogramme ⇒ Avec $E = \{x_i\}_{i=1}^N$ un échantillon taille N, si E iid, estimation converge vers la loi : $p(D = k) = |\{x_i \mid x_i^d = k\}| / N = \sum_{i=1}^N 1_{x_i^d = k} / N$ (x_i^d = jour du i-ème tweet)

Fenêtre de Parzen : Pour échantillon de taille N, R un hypercube de côté r, $V = r^d$ (d la dimension)
$\phi(x) = 1$ si $|x_i| \leq 1/2$, sinon 0 (fct. indicatrice de l'hypercube unitaire) et ϕ définie un hypercube unitaire centré à l'origine. $\phi\left(\frac{x_0 - x}{r}\right) = 1$ ssi x dans l'hypercube V centré en x_0.
Nb. échantillons dans l'hypercube : $k = \sum_{i=1}^N \phi((x_0 - x_i)/r)$ $\delta(x) = \frac{1}{V}\phi\left(\frac{x}{r}\right)$
Densité estimée ⇒ $p(x_0) = \frac{1}{N}.\sum_{i=1}^N \frac{1}{V}.\phi((x_0 - x_i)/r) = \frac{1}{N}\sum_{i=1}^N \delta(x_0 - x_i)$

KNN : k = nb. voisins à prendre en compte, $p(y|x) = \frac{1}{k}.\sum_{j, x_j \in \{k\text{-plus proches}\}} y_j$

Regression linéaire : $f_w(x) = w_0 + \sum_{i=1}^d w_i.x_i$, $f(x_i)$ doit approcher le plus y_i
Erreur ⇒ MSE = $\ell(f(x),y) = (f(x) - y)^2$. On veut minimiser $\mathbb{E}[\ell(f(x),y)]$.
Trouver $w \in \mathbb{R}^{d+1}$ minimise : $\frac{1}{n}\sum_{i=1}^n \ell(f_w(x),y) = \frac{1}{n}\sum_{i=1}^n (y^i - f(x^i))^2 = \frac{1}{n}\sum_{i=1}^n \left(y^i - w_0 - \sum_{j=1}^d w_j x_j^i\right)^2$
Fonction $L(w) : \mathbb{R}^{d+1} \to \mathbb{R}$ est convexe donc on annule son gradient.
Trouver w^* tq. $\nabla_w L(w^*) = 0$. $L(w) = (XW - Y)^T(XW - Y)$, $\nabla_w L = 2X^T(XW - Y)$
w optimal quand $\nabla_w L = 0$, solution : $(X^TX)^{-1}(X^TY)$ qui est rarement possible.

Regression logistique : $p(y = 1|x) = \mu(x)$ et $p(y = -1|x) = 1 - \mu(x)$ (cas 2D)
$p(y|x) = \mu(x)^{\frac{y+1}{2}}.(1 - \mu(x))^{\frac{1-y}{2}}$ → fct. sigmoïde : $\mu(x) = \sigma(f_w(x)) = \frac{1}{1 + e^{-f_w(x)}}$.
$\mu(x) = \sigma(f_w(x)) = \frac{1}{1 + e^{-f_w(x)}}$ et $1 - \mu(x) = \frac{1}{1 + e^{f_w(x)}} = \sigma(-f_w(x))$.
On cherche à maximiser $P(y^1, \dots, y^n | x^1, \dots, x^n) = \prod_{i=1}^n P(y^i|x^i)$
$\iff$ max. $\log(\prod_{i=1}^n P(y^i|x^i))$ $\iff$ max. $\sum_{i=1}^n \log P(y^i|x^i)$ $\iff$ min. $\sum_{i=1}^n \log\left(\frac{1}{P(y^i|x^i)}\right)$
$\iff$ min. $\sum_{i=1}^n \log(1/\sigma(-y^i.f_w(x^i)))$
On cherche $\text{argmin}_w \sum_{i=1}^n \log[1 + \exp(-y^i.f_w(x^i))] = w^*$ (par descente de gradient)

Convexe → $f'' \geq 0$ Concave → $f'' \leq 0$

Algorithme du gradient :
1) Choisir un point x_0
2) Itérer :
 - calculer $\nabla f(x_t)$
 - $x_{t+1} \leftarrow x_t - \alpha \nabla f(x_t)$

Dev. de Taylor :
$f(x) = f(x_1) + \nabla f(x_1) \times (x - x_1) + O(\|x - x_1\|^2)$
$f(x_1 + hu) - f(x_1) = h\nabla f(x_1)u + h^2 O(1)$
Minimiser $\nabla f(x_1)u$ et $u = -\frac{\nabla f(x_1)}{\|\nabla f(x_1)\|}$

Hors-ligne/batch → itère sur tous les exemples et corrections de w
Stochastique → correction par un exemple tiré aléatoirement
Batch = + stable, + rapide Stoch. = résistance au bruit

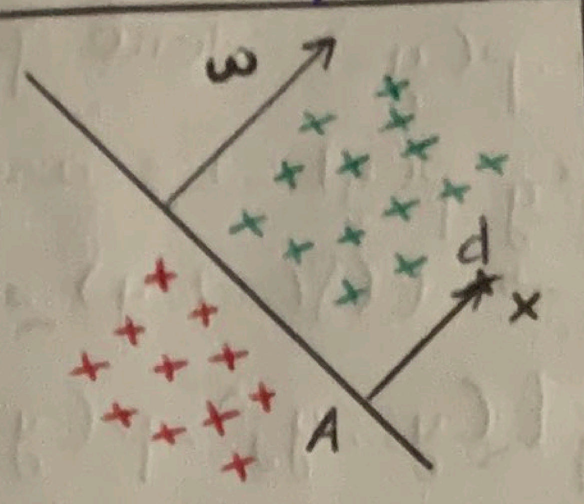
Perceptron :
Initialiser w random
Tant que pas convergence :
- $\forall (x^i, y^i)$:
  si $(y_i \times \langle w.x^i \rangle) < 0$:
  $w = w + \varepsilon y^i x^i$

Décision $= f(x) = \text{sign}(\langle w.x \rangle)$

Si $(y.\langle w.x \rangle) > 0 \rightarrow$ ne rien faire
Si $(y.\langle w.x \rangle) < 0 \rightarrow$ corriger $w = w + yx$
Hinge Loss $= \ell(f(x), y) = \max(0, \alpha - yf(x))$ avec $\alpha = 0$
Descente de gradient $\Rightarrow \ell(f_w(x), y) = \max(0, \alpha - y\langle w.x \rangle)$
$\nabla_w \ell(f_w(x), y) = \begin{cases} 0 & \text{si } (y\langle w.x \rangle) > \alpha \\ -yx_i & \text{sinon} \end{cases}$

Théorème de convergence (Novikov) $\Rightarrow$ Si $\exists R, \forall x : \|x\| \leq R$, données linéairement sép., ensemble d'apprentissage présenté assez de fois $\rightarrow$ après au plus R^2/ρ^2 corrections $\Rightarrow$ convergence.

SVM :
$d = |AX|$, $X = A + d\frac{w}{\|w\|}$
$f(A) = \langle w, A \rangle + b = 0 \Rightarrow \langle w, X - \frac{dw}{\|w\|} \rangle + b = 0$
$d = \frac{\langle w, X \rangle + b}{\|w\|} = \frac{f(x)}{\|w\|}$
$f(x) =$ fct. dist. fct. à la séparatrice



Si données séparables :
Introduction à la contrainte de distance fct. aux points les plus proches est fixée à $1 \Rightarrow f(x) = Wx_1 + b \underset{= \pm}{1}$
Points sont tq. $yf(x) \geq 1$ et γ la distance euclidienne : $x_1 - x_2 = \gamma \frac{w}{\|w\|}$
γ = distance hyperplan (frontière) et point x_1 le plus proche, γ = la marge
$\gamma \frac{ww}{\|w\|} = 1 \iff \gamma \|w\| = 1 \iff \gamma = \frac{1}{\|w\|}$ donc maximiser la marge $\iff$ minimiser $\|w\|$
Minimiser $\|w\|^2$ tq. $\forall i, (wx^i + b)y^i \geq 1$

Données bruitées SVM :
Introduction variables ressorts $\xi^i \rightarrow$ tolère débordement $(wx^i + b)y^i \geq 1 - \xi^i$, $\xi_i \geq 0$.
Miniser $\|w\|^2 + K\sum \xi_i$ tq. $(wx^i + b)y^i \geq 1 - \xi_i$ et $\xi_i \geq 0$.
$\begin{cases} \xi_i = 0 & \text{si } (wx^i + b)y^i \geq 1 \\ \xi_i = 1 - (wx^i + b)y^i & \text{si } (wx^i + b)y^i < 1 \end{cases} \Rightarrow \xi_i = \max(0, 1 - (wx^i + b)y^i)$ = hinge loss
Constante $K = C$ pour SVM. Minimiser $\|w\|^2 + K\sum \ell(y^i, wx^i + b)$ avec $\ell(y, \hat{y}) = \max(0, 1 - y\hat{y})$
$\|w\|^2 \rightarrow$ terme de régularisation pour contrôler le sur-apprentissage.

Lagrangien : Fonction auxiliaire : $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$ dont on cherche l'optimum
On cherche $\nabla \mathcal{L}_{x,\lambda}(x, \lambda) = 0$ soit : $\frac{d\mathcal{L}}{d\lambda} = 0 = g(x)$ et $\nabla_x \mathcal{L}(x, \lambda) = 0$, $\nabla_x f(x) = \lambda \nabla_x g(x)$
Formulation duale : contrainte d'inégalité $c_i \rightarrow$ introduit une variable $\lambda_i \geq 0$
contrainte d'égalité $g_j \rightarrow$ introduit une variable $\mu_j \in \mathbb{R}$
Formulation duale $= \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_i \lambda_i c_i(x) + \sum_j \mu_j g_j(x)$
$\min_x f(x)$ tq. $c_i(x) \leq 0$, $g_j(x) = 0 \iff \min_x \max_{\lambda,\mu} \mathcal{L}(x, \lambda, \mu)$

Cas simple (sans slack) :
$\min_{w,b} \frac{1}{2}\|w\|^2$ tq $y_i(wx^i + b) \geq 1$
$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i (y_i(wx_i + b) - 1)$
$\nabla_w L = w - \sum_i \alpha_i y_i x_i = 0$
$\iff w = \sum_i \alpha_i y_i x_i$
$\nabla_b L = \sum_i \alpha_i y_i = 0$
$\Rightarrow \max -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$
tq. $\sum \alpha_i y^i = 0$ et $\alpha_i \geq 0$

Cas compliqué (avec slack) :
$\min_{w,b} \frac{1}{2}\|w\|^2 + K\sum \xi_i$ tq. $\xi_i \geq 0$ et $y^i(wx^i + b) \geq 1 - \xi_i$
$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + K\sum_i \xi_i - \sum_i \beta_i \xi_i - \sum_i \alpha_i (y^i(wx^i + b) + \xi_i - 1)$
$\nabla_w L(w, b, \alpha, \xi, \beta) = w - \sum_i \alpha_i y^i x^i = 0$
$\nabla_b L = \sum_i \alpha_i y_i = 0$
$\nabla_\xi L = K - \alpha_i - \beta_i = 0$
$\Rightarrow \max_\alpha -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$
tq. $\sum_i \alpha_i y^i = 0$ et $\alpha_i \in [0, K]$
Conditions d'optimalité :
$\begin{cases} \alpha_i(y^i(wx^i + b) + \xi_i - 1) = 0 \\ \beta_i \xi_i = 0 \end{cases}$
$\begin{cases} \alpha_i = 0 \Rightarrow y^i(wx^i + b) \geq 1 \\ 0 < \alpha_i < K \Rightarrow (y^i(wx^i + b) - 1) = 1 \\ \alpha_i = K \Rightarrow (y^i(wx^i + b) - 1) \leq 1 \end{cases}$