PSL

Apprentissage Supervisé

Classification de textes

Jean Barré

01 Mars 2025

ENS-PSL

Apprentissage Supervisé vs Non Supervisé

Apprentissage Supervisé vs Non Supervisé

- Apprentissage Supervisé :
  - Le modèle est entraîné sur des données étiquetées, c'est-à-dire avec des réponses connues.
  - Exemple : classifier des romans en fonction de leur sous-genres romanesques

- Apprentissage Non Supervisé :
  - Le modèle apprend sans réponses connues, en identifiant lui-même des patterns dans les données.
  - Exemple : identifier des groupes de mots fréquents dans un corpus de textes.

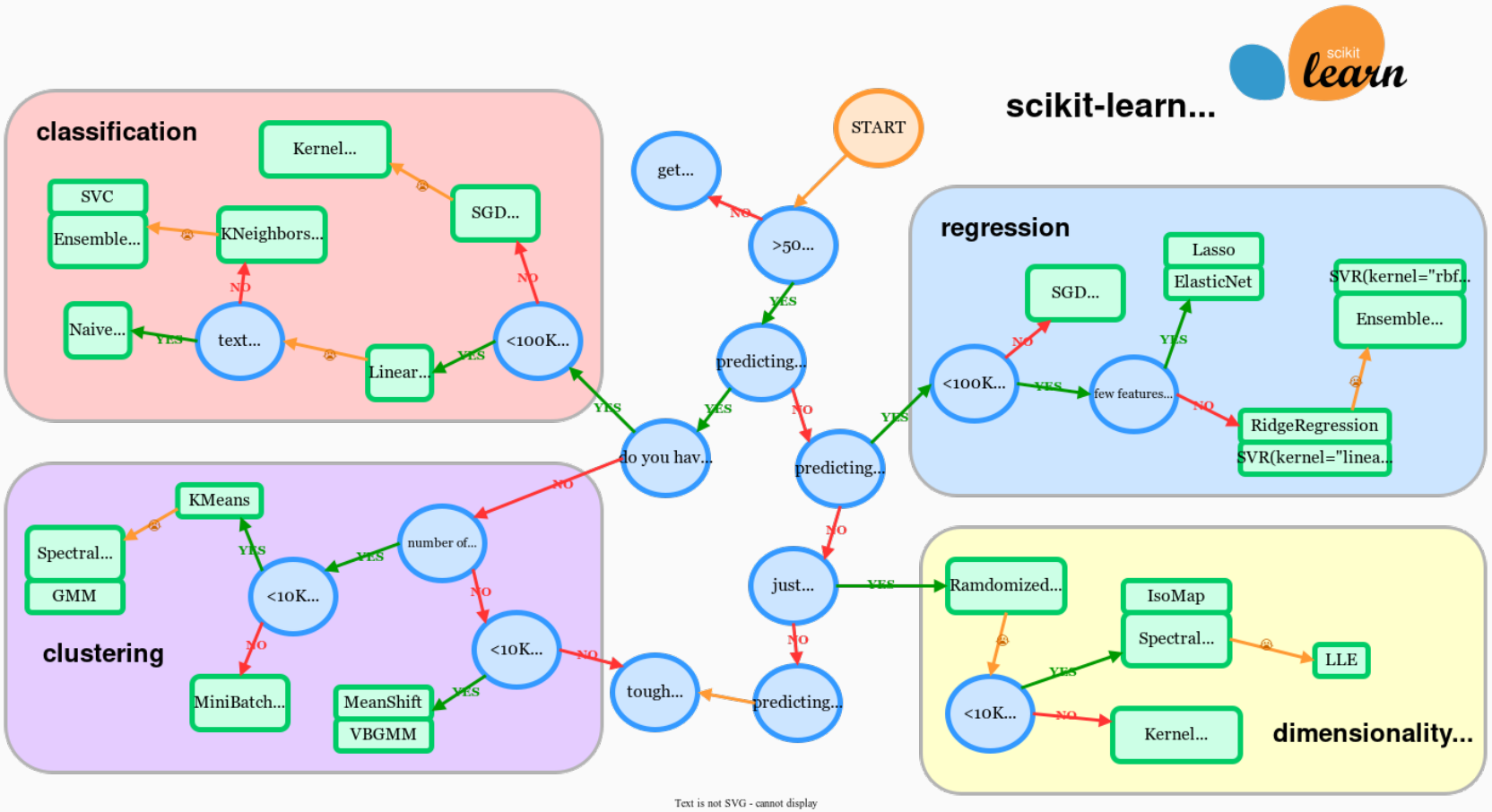
Classification vs Régression

- **Classification** :
  - Le modèle attribue des catégories ou classes aux données.
  - Exemple : déterminer si un texte est de type « fiction » ou « essai ».

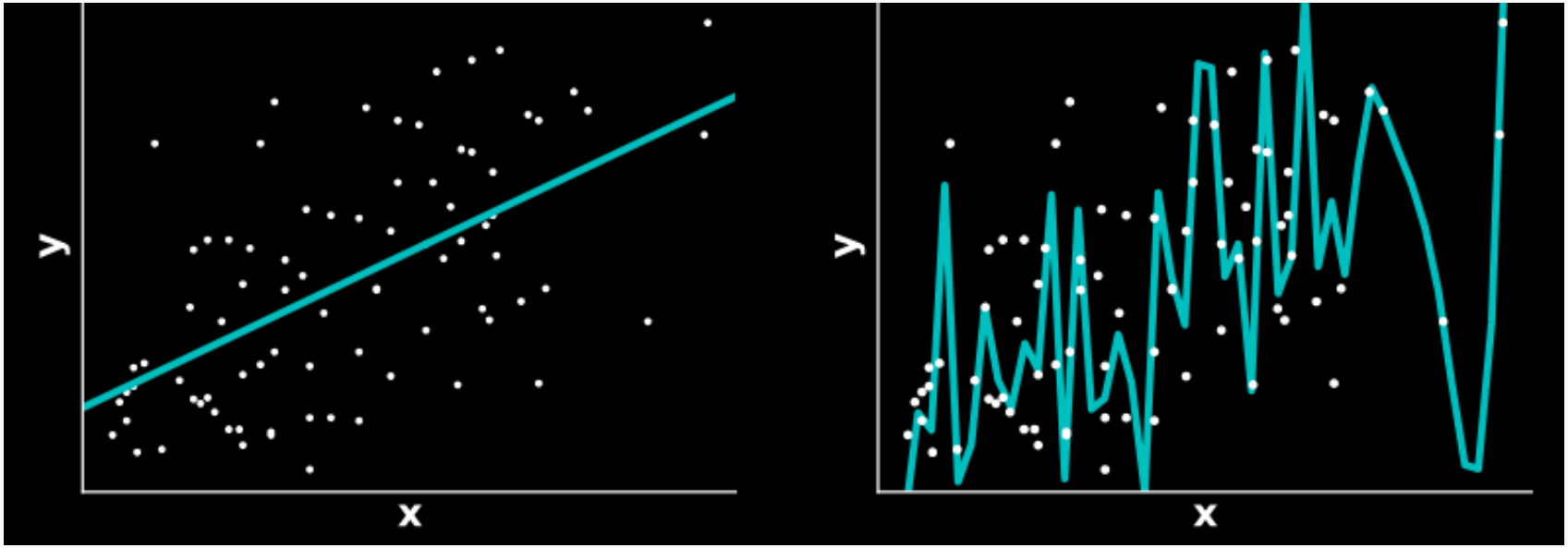
- **Régression** :
  - Le modèle prédit des valeurs continues.
  - Exemple : estimer la date approximative d'un texte à partir de caractéristiques textuelles.

Scikitlearn

version SVG



Surentraînement vs Sous-entraînement

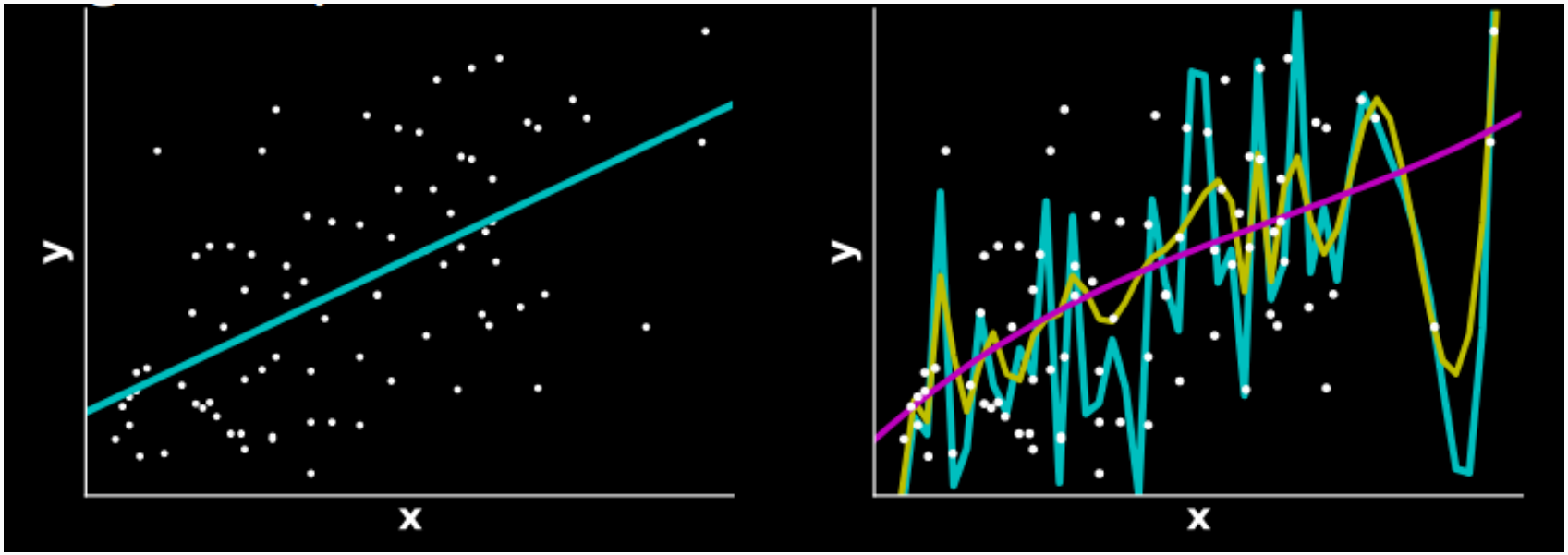


Surentraînement vs Sous-entraînement

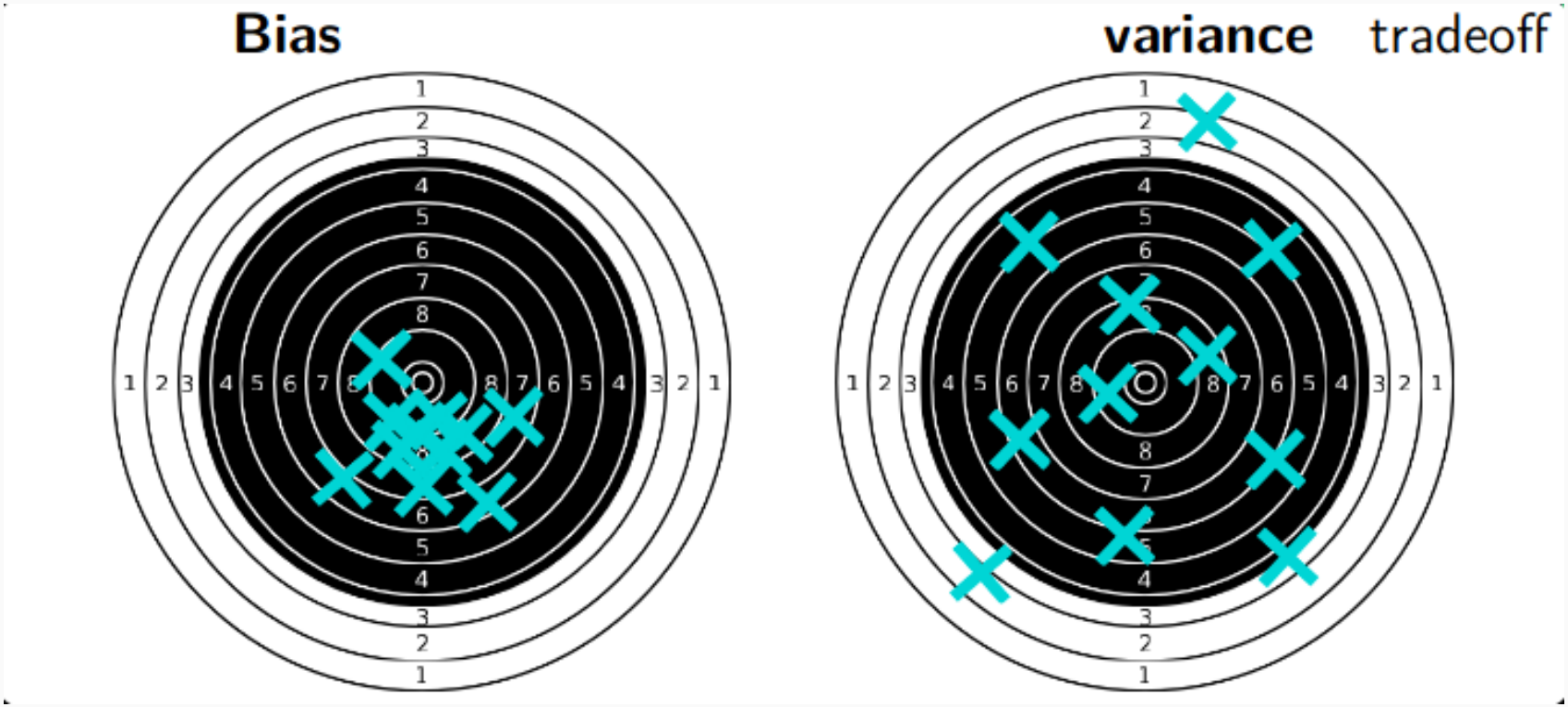
- Sous-entraînement :
  - Le modèle est trop simple et n'arrive pas à capturer les patterns des données.
  - Se traduit par des erreurs élevées tant sur les données d'entraînement que de test.

- Surentraînement :
  - Le modèle est trop complexe et s'adapte excessivement aux données d'entraînement.
  - Il capture le bruit et les variations spécifiques, ce qui nuit à sa capacité de généralisation.

Régularisation



Généralisation



Mémoriser vs Généraliser

- **Mémoriser** :
  - Le modèle retient les exemples spécifiques vus pendant l'apprentissage.
  - Risque de mauvaise performance sur de nouveaux exemples.

- **Généraliser** :
  - Le modèle extrait des patterns pertinents qui s'appliquent à des données inédites.
  - Assure une bonne performance sur des cas non vus.

Objectifs Principaux du Machine Learning

- Extraire les patterns de la matrice de données X.

- Déduire une structure capable de prédire ou d'estimer une cible Y (par exemple : le genre d'un texte, l'époque d'un document).

La modélisation

- **Modélisation Descriptive** :
  - Comprendre et résumer les informations contenues dans les données.
  - Exemple : analyser des tendances historiques ou identifier des groupes similaires.

- **Modélisation Prédictive** :
  - Utiliser les patterns pour faire des prédictions sur de nouvelles données.
  - Exemple : prédire l'auteur probable d'un texte non attribué.

Données d'Entraînement vs Données de Test

- **Données d'Entraînement** :
  - Utilisées pour apprendre et entraîner le modèle.

- **Données de Test** :
  - Utilisées pour évaluer la performance du modèle sur des exemples non vus durant l'entraînement.

Classification

Définition

- Une fonction de mappage / correspondance h qui associe une donnée $x \in \mathcal{X}$ à une étiquette $y \in \mathcal{Y}$ l'ensemble de sortie possible.

Exemple

- $\mathcal{X}$ = ensemble de tous les documents.
- $\mathcal{Y} = \{$anglais, mandarin, grec, ...$\}$.
  Avec
- $x =$ un document unique.
- $y =$ grec ancien.
- $h(x) = y$, par exemple :

$$h(\text{μῆνιν ἄειδε θεὰ}) = \text{grec ancien.}$$

Approximation de la fonction de classification

Recherche de la meilleure approximation

- Soit $h(x)$ la fonction « vraie ». Nous ne la connaissons jamais.
- Comment trouver la meilleure approximation $\hat{h}(x)$ de $h(x)$?

Exemple de règle basée

- Si x contient des caractères dans la plage Unicode `0370-03FF` : alors $\hat{h}(x) = \text{grec}$.

Apprentissage supervisé

Principe

- À partir de données d'entraînement sous forme de paires $\langle x, y \rangle$, on apprend une fonction $\hat{h}(x)$ qui approxime $h(x)$.

Exemples de tâches de classification

- Identification de la langue : texte {*anglais*, *mandarin*, *grec*, ...}.
- Classification de spam : email {*spam*, *non-spam*}.
- Attribution d'auteur : texte {*JK Rowling*, *James Joyce*, ...}.
- Classification par genre : roman {*policier*, *romance*, *gothique*, ...}.
- Analyse des sentiments : texte {*positif*, *négatif*, *neutre*, *mixte*}.

Analyse de sentiment

Analyse de sentiment au niveau du document

- Déterminer si le texte complet est positif ou négatif (ou les deux/neutre) par rapport à une cible implicite.
- Exemple : critiques de films (Pang et al. 2002, Turney 2002).

Données d'entraînement

Exemples de textes annotés

- **Exemple négatif** :

  « J'ai détesté ce film. Détesté, détesté, détesté, détesté, détesté ce film. Je l'ai détesté. J'ai détesté chaque instant ridicule et insultant pour le public. J'ai détesté la sensibilité qui pensait que quelqu'un pourrait l'apprécier. »

- **Exemple positif** :

  « … est un film qui provoque de vrais frissons, non figuratifs, le long de ma colonne vertébrale, et c'est sans doute le fruit le plus courageux et ambitieux du génie de Coppola. »

- Sources : Roger Ebert, *Apocalypse Now* (positif) et *North* (négatif).

Sentiment comme tonalité

Définition

- Le sentiment ne mesure plus l'attitude du locuteur vis-à-vis d'une cible particulière, mais la tonalité générale (positive ou négative) qui s'en dégage.

Sentiment comme tonalité (suite)

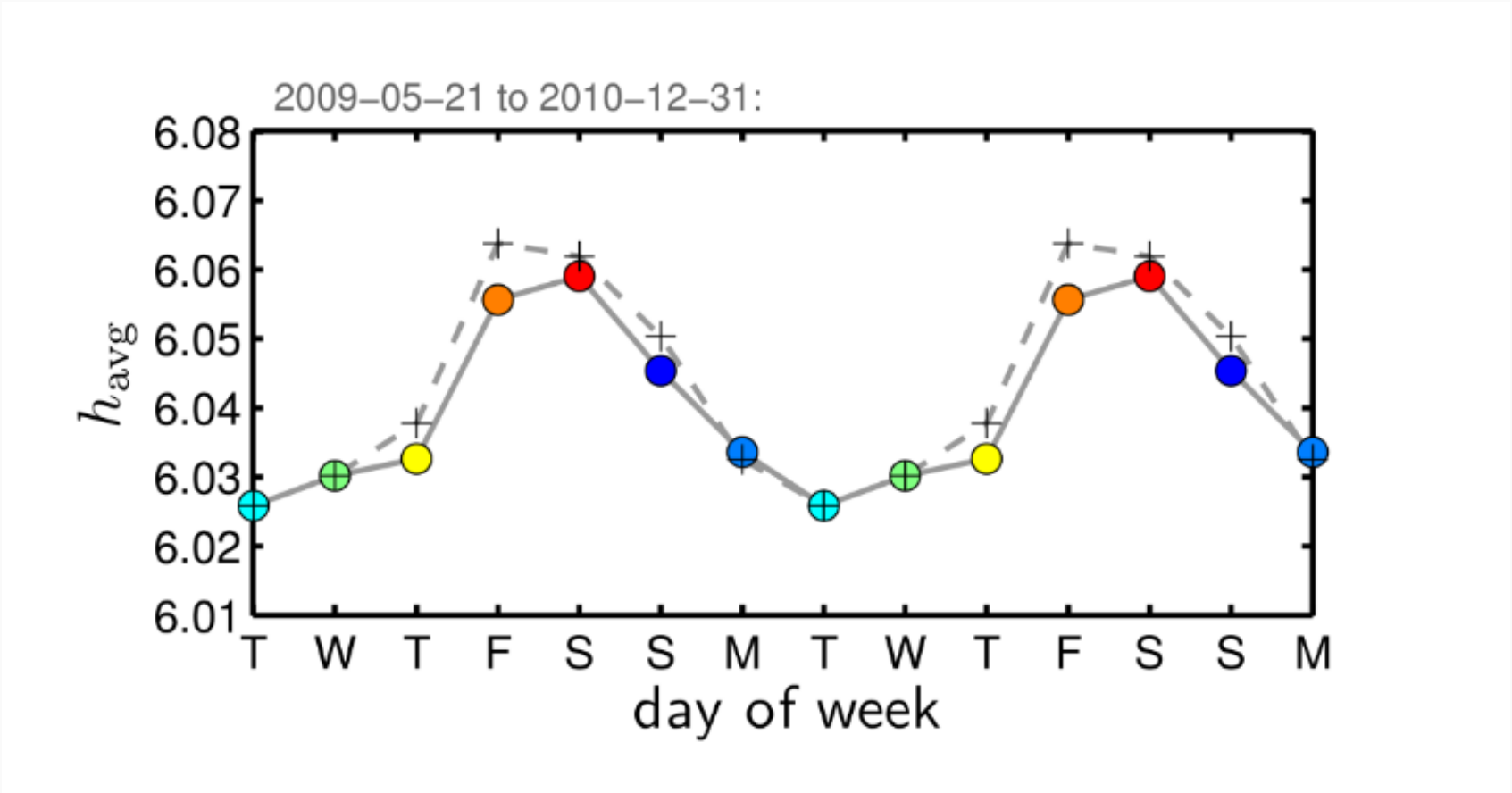


Fig. 1 : Dodds et al. (2011) : « Temporal patterns of happiness and information in a global social network : Hedonometrics and Twitter ».

Sentiment comme tonalité (suite)

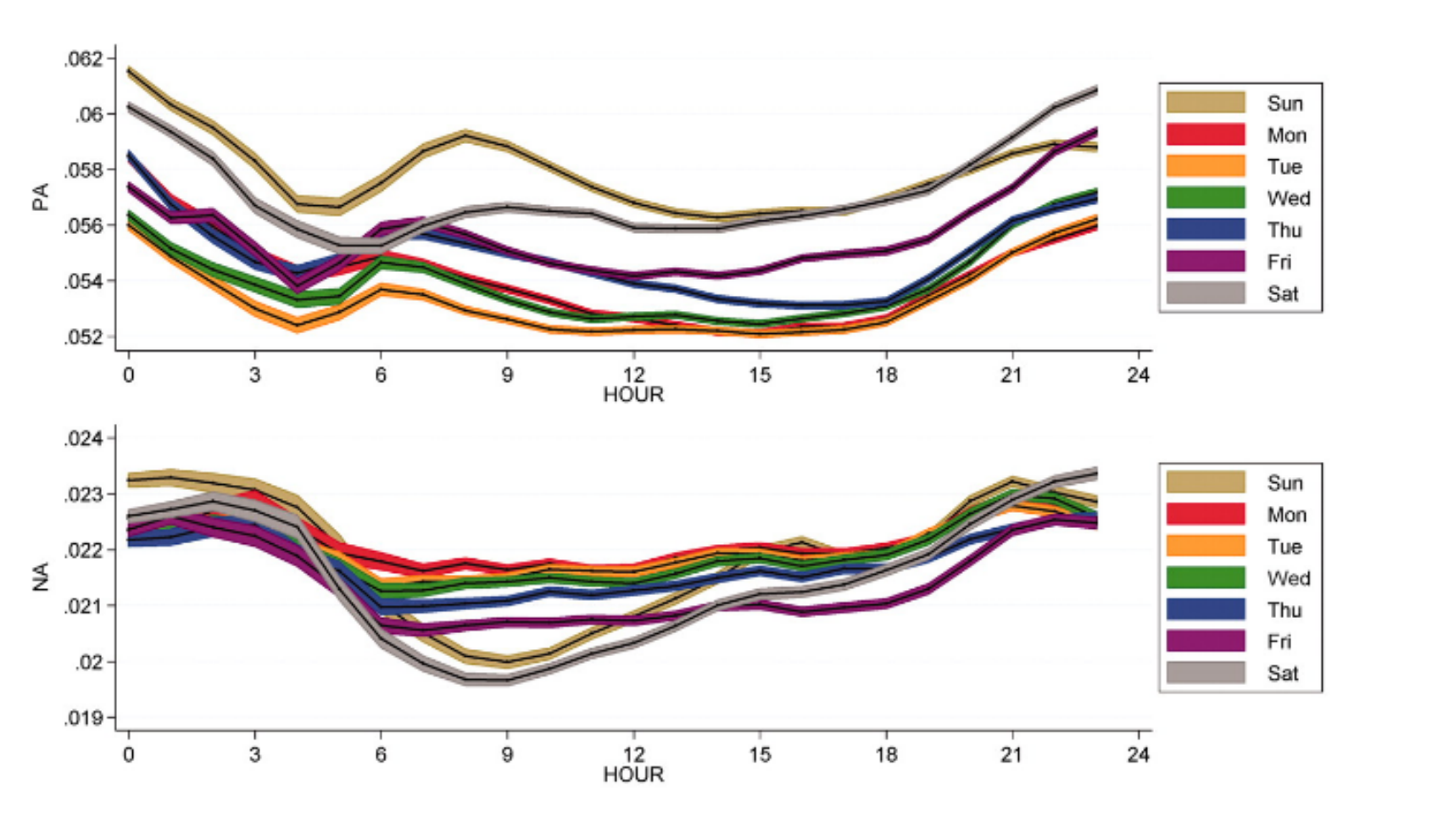


Fig. 2 : Golder et Macy (2011) : « Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures »

Dictionnaires de sentiment

- General Inquirer (1966).
- MPQA subjectivity lexicon (Wilson et al. 2005) `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`.
- LIWC (Linguistic Inquiry and Word Count, Pennebaker 2015). -> Mesure de l'affect positif (PA) et de l'affect négatif (NA)
- AFINN (Nielsen 2011).
- NRC Word-Emotion Association Lexicon (EmoLex), Mohammad et Turney (2013).

LIWC (Linguistic Inquiry and Word Count)

- Comprend 73 lexiques distincts, conçus pour des applications en psychologie sociale.

Positive Emotion	Negative Emotion	Insight	Inhibition	Family	Negate
appreciat*	anger*	aware*	avoid*	brother*	aren't
comfort*	bore*	believe	careful*	cousin*	cannot
great	cry	decid*	hesitat*	daughter*	didn't
happy	despair*	feel	limit*	family	neither
interest	fail*	figur*	oppos*	father*	never
joy*	fear	know	prevent*	grandf*	no
perfect*	griev*	knew	reluctan*	grandm*	nobod*
please*	hate*	means	safe*	husband	none
safe*	panic*	notice*	stop	mom	nor
terrific	suffers	recogni*	stubborn*	mother	nothing
value	terrify	sense	wait	niece*	nowhere
wow*	violent*	think	wary	wife	without

Pourquoi l'analyse de sentiment est-elle difficile ?

- Le sentiment mesure l'état privé d'un locuteur, qui est inobservable.
- Parfois, des mots simples (ex. : *love*, *amazing*, *hate*, *terrible*) sont de bons indicateurs, mais souvent une connaissance approfondie du contexte est nécessaire.
- Exemple :

  > *« Valentine's Day is being marketed as a Date Movie. I think it's more of a First-Date Movie. If your date likes it, do not date that person again. And if you like it, there may not be a second date. »*
  >
  > — Roger Ebert, *Valentine's Day*

Exemples de classification de sentiment

Principe

- À partir de données d'entraînement sous forme de paires $\langle x, y \rangle$, on apprend une fonction $\hat{h}(x)$ qui approxime la véritable fonction $h(x)$.

- « loved it ! » $\rightarrow$ positif.
- « terrible movie » $\rightarrow$ négatif.
- « not too shabby » $\rightarrow$ positif.

Classification par apprentissage supervisé

Composantes de $\hat{h}(x)$

- La structure formelle de la méthode d'apprentissage (ex. : Naive Bayes, régression logistique, réseau de neurones convolutionnel, etc.).
- La représentation des données.

Représentation pour l'analyse de sentiment

- Utilisation uniquement des mots positifs ou négatifs issus du LIWC.
- Représentation en sac de mots (mots isolés).
- Conjonctions de mots (ngrams séquentiels, skip-ngrams, autres combinaisons non linéaires).
- Structures linguistiques de haut niveau (ex. : syntaxe).

Représentation par sac de mots

Mot	Apocalypse Now	North
the	1	1
of	0	0
hate	0	9
genius	1	0
bravest	1	0
stupid	0	1
like	0	1

Tab. 1 : Représentation des textes par le sac de mots

Régression logistique binaire

- Espace de sortie : $Y = \{0, 1\}$.
- Modèle :

$$P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{F} x_i \beta_i\right)}$$

Régression logistique binaire

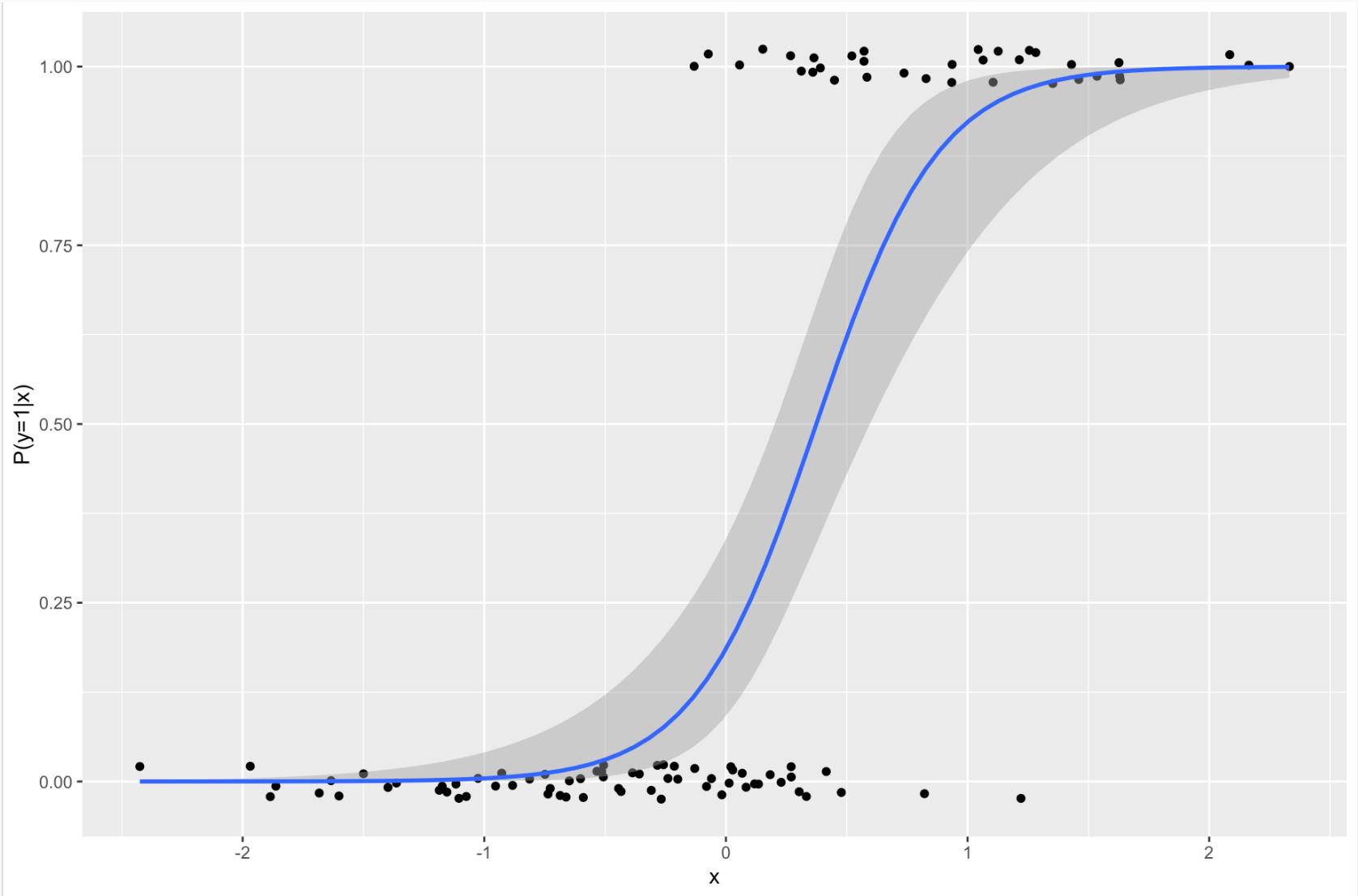


Fig. 3 : Régression Logistique

Vecteur de caractéristiques et coefficients

Exemple de caractéristiques extraites

Caractéristiques (Valeurs)

Mot	Valeur
the	0
and	0
bravest	0
love	0
loved	0
genius	0
not	0
fruit	1
BIAS	1

Coefficients (β)

Mot	β
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
not	-3.0
fruit	-0.8
BIAS	-0.1

Exemple illustratif de vecteur de caractéristiques et des coefficients pour la régression logistique.

Régression logistique

- La régression logistique est un classificateur discriminatif.
- Elle n'assume pas l'indépendance des caractéristiques.
- Sa force réside dans la possibilité d'utiliser des caractéristiques riches et expressives sans la contrainte d'indépendance.

Caractéristiques de la tâche

Les caractéristiques sont l'endroit où vous pouvez encoder votre propre compréhension du problème du problème.

- **Unigrams** : ex. *like*.
- **Bigrams** : ex. *not like*, et n-grammes d'ordre supérieur.
- **Préfixes** : mots commençant par, par exemple, *in-*.
- **Présence dans un dictionnaire de sentiment positif** : indique si le mot figure dans une liste prédéfinie.

Vecteur de caractéristiques et coefficients

Exemple de caractéristiques extraites

Comment avoir de bonnes valeurs de β ?

Coefficients (β)

Mot	β
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
not	-3.0
fruit	-0.8
BIAS	-0.1

Vraisemblance conditionnelle

- Pour l'ensemble des données d'entraînement, on souhaite que la probabilité du label vrai y pour chaque exemple x soit élevée.
- La vraisemblance conditionnelle s'exprime par :

$$\prod_{i=1}^{N} P(y_i \mid x_i, \beta)$$

- L'objectif est de choisir les paramètres β qui maximisent cette probabilité.

Exemple numérique

- Considérons un exemple avec des caractéristiques telles que `love` et `loved` :

Exemple	`love`	`loved`	$a = \sum x_i \beta_i$	$\exp(-a)$	$P(y = 1 \mid x, \beta)$
x_1	1	0	3.0	0.05	95.2%
x_2	1	1	4.2	0.015	98.5%
x_3	0	0	-0.1	1.11	41.5%

Maximisation de la log-vraisemblance

- Maximiser la vraisemblance revient à maximiser la log-vraisemblance :

$$\arg\max_{\beta} \prod_{i=1}^{N} P(y_i \mid x_i, \beta) \;=\; \arg\max_{\beta} \sum_{i=1}^{N} \log P(y_i \mid x_i, \beta)$$

- On définit alors la fonction objective :

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \log P(y_i \mid x_i, \beta)$$

- La maximisation de $\mathcal{L}(\beta)$ permet d'obtenir les meilleurs paramètres.

Descente de gradient

- Pour maximiser $\mathcal{L}(\beta)$, on utilise la descente de gradient.
- Le gradient de la log-vraisemblance est donné par :

$$\nabla \mathcal{L}(\beta) = \sum_{\langle x,y \rangle} (y - \hat{p}(x))\, x_i$$

- **Exemples :**
  - Si $y = 1$ et $\hat{p}(x) = 0.99$, la mise à jour des poids sera faible.
  - Si $y = 1$ et $\hat{p}(x) = 0$, la mise à jour sera importante.
- La *descente de gradient stochastique* met à jour β après chaque point de données.

Optimisation pratique

- Lors du calcul de $P(y \mid x)$ ou du gradient, il suffit de considérer les caractéristiques non nulles.
- Ceci rend particulièrement utile l'utilisation de valeurs binaires et d'une représentation parcimonieuse.
- La régression logistique s'exprime ainsi :

$$P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{F} x_i \beta_i\right)}$$

Sélection de caractéristiques et régularisation

- De nombreuses caractéristiques apparaissent rarement, parfois par pur hasard.
- Deux approches possibles :
  - Seuil par fréquence minimale (ce qui peut éliminer des informations utiles).
  - Intégrer une croyance a priori que tous les β doivent être nuls sauf en présence d'une forte évidence.

Régularisation L2

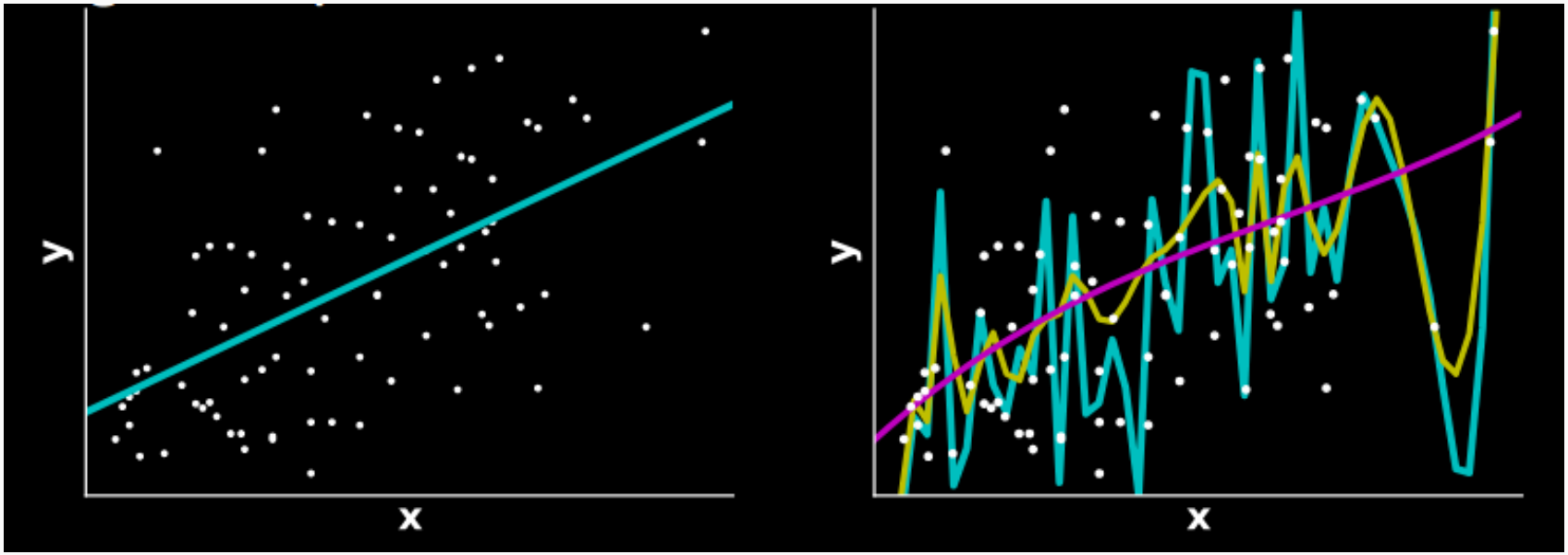
- On ajoute une pénalité quadratique aux valeurs élevées de β :

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \log P(y_i \mid x_i, \beta) - \eta \sum_{j=1}^{F} \beta_j^2$$

- Ceci équivaut à supposer que chaque β_j provient d'une distribution normale centrée sur 0.
- Le paramètre η contrôle la force de la pénalité (optimisé sur les données d'entraînement).

Exemple de régularisation L2

- Sans régularisation, certains coefficients peuvent être très élevés.
- Avec une régularisation modérée, les coefficients diminuent.
- Une forte régularisation pousse les coefficients vers zéro.



Régression logistique multiclasse

- Pour une classification à K classes, le modèle s'exprime comme suit :

$$P(Y = y \mid X = x; \beta) = \frac{\exp(x \cdot \beta_y)}{\sum_{y' \in Y} \exp(x \cdot \beta_{y'})}$$

- Ici, $Y = \{1, \ldots, K\}$ et chaque classe dispose de son propre vecteur de coefficients β_y.

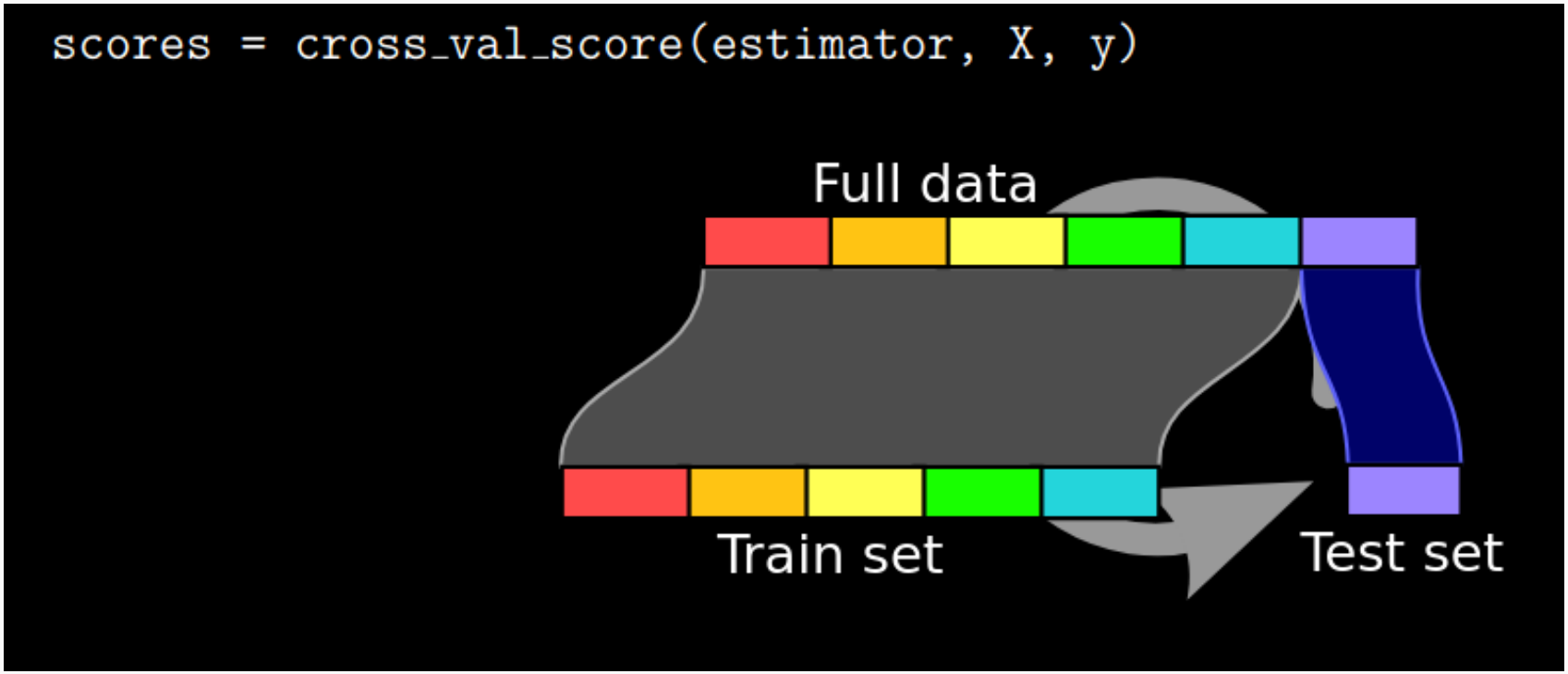
Exemple de coefficients en multiclasse

Caractéristique	Positive	Négative	Neutre
the	1.33	-0.80	-0.54
and	1.21	-1.73	-1.57
bravest	0.96	-0.05	0.24
love	1.49	0.53	1.01
loved	-0.52	-0.02	2.21
genius	0.98	0.77	1.53
not	-0.96	2.14	-0.71
fruit	0.59	-0.76	0.93
BIAS	-1.92	-0.70	0.94

- Note : Ici, trois ensembles de coefficients sont utilisés, un pour chaque classe (positive, négative, neutre).

Évaluation

- Une partie critique dans le développement de nouveaux algorithmes et méthodes est l'évaluation, qui permet de démontrer leur efficacité.



Conception de l'expérience

Données et partitionnement

- Espace d'instance : $\mathcal{X}$.

- Division typique des données :
  - **Entraînement** (80%) : apprentissage des modèles.
  - **Développement** (10%) : sélection du modèle.
  - **Test** (10%) : évaluation finale (à ne jamais consulter avant la fin).

Matrice de confusion multi-classe

Exemple de matrice de confusion

	Positive	Negative	Neutral
Positive	100	2	15
Negative	0	104	30
Neutral	30	40	70

Lignes : étiquettes réelles ; Colonnes : étiquettes prédites ($\hat{y}$).

Exactitude (Accuracy)

Définition

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I[\hat{y}_i = y_i],$$

où $I[x]$ vaut 1 si la condition est vraie, 0 sinon.

Précision

Définition (pour la classe Positive)

$$\text{Précision}(\text{POS}) = \frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{POS})}{\sum_{i=1}^{N} I(\hat{y}_i = \text{POS})}.$$

La précision représente la proportion d'éléments prédits comme appartenant à une classe qui le sont réellement.

Rappel

Définition (pour la classe Positive)

$$\text{Rappel(POS)} = \frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{POS})}{\sum_{i=1}^{N} I(y_i = \text{POS})}.$$

Le rappel mesure la proportion des instances réelles d'une classe correctement prédites.

F-score

Définition

$$F = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}.$$

Le F-score est la moyenne harmonique entre la précision et le rappel.

Baseline de la classe majoritaire

- Choisir l'étiquette la plus fréquente dans les données d'entraînement (ne pas tenir compte des données de test).

- Prédire cette étiquette pour chaque point de données du test.