

# Statistiques pour l'évaluation

Damien Nouvel



# Plan

1. Pourquoi évaluer
2. Mesures ensemblistes
3. Adaptation aux tâches
4. Evaluation de paramètres à partir des données

# Tâche et données

- ▶ Mesurer la **performance** pour une tâche donnée
- ▶ L'**évaluation** quantitative en TAL met en jeu
  - Une **tâche** à accomplir automatiquement
  - Un jeu de **données** pour
    - **Entraîner** (paramétrer) les modèles
    - **Développer** (améliorer) des modèles
    - **Évaluer** les modèles
- ⇒ **Efforts** pour préparer les données
- ⇒ Éviter les **biais**
- ▶ **Corrélations** entre **modélisation** et **évaluation**
  - Pour une tâche, les **performances** varient fortement selon
    - **Types** de données (domaine, modalité, qualité, etc.)
    - **Mesure** d'évaluation utilisée
    - ...
- ⇒ Bien tenir compte des **conditions** de l'évaluation
- ⇒ Recherche d'**objectivité** et de **fiabilité**

# Plan

1. Pourquoi évaluer
2. Mesures ensemblistes
3. Adaptation aux tâches
4. Evaluation de paramètres à partir des données

# Mesures binaires

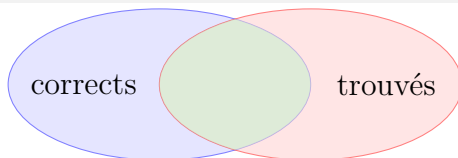
- ▶ Sur  $N$  données, calcul d'un **taux d'erreur** binaire
- ▶ Exactitude (en anglais **accuracy**)
- ▶ Exemple : “Je suis parti en vacances avec Marie” (7 mots)
  - Combien de données ont été **correctement traitées** ?
  - ⇒ Etiquetage morpho-syntaxique
    - PRO VER VER PREP NC PREP NP :  $7/7 = 100\%$
    - PRO VER **NC** PREP NC PREP NP :  $6/7 = 86\%$
  - ⇒ Extraction de noms propres
    - NA NA NA NA NA NA NP :  $7/7 = 100\%$
    - NA NA NA NA NA NA **NA** :  $6/7 = 86\%$
- ▶ Terminologie de l'évaluation :
  - Tests statistiques (maladies) : positif, négatif, vrai, faux
  - Recherche d'information (documents) : trouvé, correct

# Table de contingence

	<i>malade</i>	<i>non-malade</i>	<i>totaux</i>
<i>positifs</i>	30 (VP)	10 (FP)	40
<i>négatifs</i>	20 (FN)	50 (VN)	70
<i>totaux</i>	50	60	110

- ▶ Valeurs extraites de la table
  - **Vrais** (VP, VN) : test en accord avec la maladie
  - **Faux** (FP, FN) : test en désaccord avec la maladie
- ▶ Mesures pour **évaluer** les tests (rappels)
  - **Sensibilité** :  $\frac{|VP|}{|VP| + |FN|} = \frac{|VP|}{|malade|} = \frac{30}{50} = 60\%$
  - ⇒ Taux de détection de la **présence** de maladie ?
  - **Spécificité** :  $\frac{|VN|}{|VN| + |FP|} = \frac{|VN|}{|non - malade|} = \frac{50}{60} = 83\%$
  - ⇒ Taux de détection de l'**absence** de maladie ?

# Précision, rappel, f-mesure



- ▶ Autre terminologie
  - **Trouvés** (positifs), **corrects** (malades)
  - ⇒ Pas d'équivalent pour **vrais** ou **faux**
- ▶ Mesures pour **évaluer** la **recherche d'information**

- **Précision** :  $P = \frac{|corrects \wedge trouvés|}{|trouvés|}$  (et **bruit**)

- **Rappel** :  $R = \frac{|corrects \wedge trouvés|}{|corrects|}$  (et **silence**)

- **F-mesure** :  $F_1 = \frac{2 * P * R}{P + R}$

⇒ Par exemple, 10 premiers résultats d'un moteur de recherche

# Moyennes de mesures

⇒ Lorsque l'on mesure **plusieurs dimensions** (types)

<i>type</i>	<i>trouvés</i>	<i>corrects</i>	<i>trouvés et corrects</i>	<b>P</b>	<b>R</b>
personnes	90	100	80	0.89	0.8
organisations	15	20	10	0.67	0.5

## ▸ Agrégation de résultats

### • Macro-moyenne

- Calcul des mesures, puis moyennes

- Précision :  $\frac{0.89 + 0.67}{2} = 0.78$  / Rappel :  $\frac{0.8 + 0.5}{2} = 0.65$

⇒ Chaque type a un poids égal dans la mesure finale

### • Micro-moyenne

- Sommes sur les valeurs les plus fines, puis calcul de la mesure

- Précision :  $\frac{80 + 10}{90 + 15} = 0.86$  / Rappel :  $\frac{80 + 10}{100 + 20} = 0.75$

⇒ Poids des types selon leurs nombres d'éléments



# Plan

1. Pourquoi évaluer
2. Mesures ensemblistes
3. Adaptation aux tâches
4. Evaluation de paramètres à partir des données

# Séquences textuelles (énoncés)

- ▶ Reconnaissance de parole / écriture
  - Erreurs par mots (**insertions, substitutions**)  
⇒ **WER** (Word Error Rate)
- ▶ Reconnaissance d'entités nommées (**types**)
  - Classification de segments (**types et extensions**)  
⇒ **SER** (Slot Error Rate)
- ▶ Traduction (**présence de mots traduits**)
  - Qualité globale du texte, dont vocabulaire  
⇒ **BLEU** (Bilingual Evaluation Understudy)
- ▶ Résumé automatique
  - Qualité globale du texte, dont vocabulaire
  - **ROUGE** (Recall-Oriented Understudy of Gisting Eval.)

# Données ordonnées

- ▶ Nombreux résultats (moteurs de recherche)
  - **Pertinence** des premiers résultats
  - Moins d'importance en **fin de classement**
- ▶ Mesures selon les  $k$  premiers résultats sur  $N$ 
  - Précision au rang  $k$  :  $P(k)$  (et rappel  $R(k)$ )
  - R-Précision, selon le nombre de résultats attendus
  - Précision moyenne :

$$MoyP = \sum_{k=1}^N P(k) * (R(k) - R(k-1))$$

- nDCG : normalized discounted cumulative gain

⇒ Prise en compte de l'ordre en **recherche d'information**

# Annotation de corpus

- ▶ Accord inter-annotateur

# Plan

1. Pourquoi évaluer
2. Mesures ensemblistes
3. Adaptation aux tâches
4. Evaluation de paramètres à partir des données

# Jeux de données

- ▶ Pour un jeu de données
  - **Entraîner, développer, évaluer** le(s) modèle(s)
  - Attention au **sur-apprentissage**!

⇒ La connaissance du jeu de données **biaise** la modélisation

⇒ Par exemple, apprentissage **par cœur**
- ▶ Jeux de données pour **apprentissage supervisé**
  - **Entraînement** (train) : stat. pour les modèles (probabilités)
  - **Développement** (dev) : amélioration itérative des modèles
  - **Evaluation** (test) : test final (idéalement unique)

⇒ Utilisation de **validation croisée** (cross-validation)

  - Diviser le jeu de données  $D$  en  $n$  sous-ensembles (souvent 10)
  - Pour chaque sous-ensemble  $i$  :
    - Paramétrer le modèle sur  $D - d_i$
    - Évaluer le modèle sur  $d_i$
- ▶ Méthodes non-supervisées : pas de données d'entraînement !

# Échantillonnage de données

- ▶ Pour des variables continues
  - Sélection à intervalles réguliers
  - Génération de valeurs par paramètres (moyenne, écart-type)
  - ...
- ▶ Échantillonnage d'un ensemble **très volumineux**
  - **Tirage aléatoire** des exemples
  - Sélection d'exemples **représentatifs** (similarité)
  - **Pondération** des exemples selon leur distance au centroïdes
  - ...
- ▶ Échantillonnage d'ensembles **infinis** (“Reservoir Sampling”)
  - Sélection des  $k$  premiers éléments dans le réservoir
  - Pour le autres éléments  $i > k$  :
    - Tirage au sort d'un chiffre  $0 < p < i$
    - Si  $p < k$  remplacer l'élément  $p$  du réservoir

# Généralisation vs spécialisation

- ▶ Tendances générales
  - Généralisation
    - Prise en compte de données plus diverses
    - Adaptation aux nouvelles données
    - Perte en précision
    - Gain en rappel
    - ⇒ Sous-apprentissage du modèle
  - Spécialisation
    - Focalisation sur des données spécifiques
    - Difficulté face aux nouvelles données
    - Gain en précision
    - Perte en rappel
    - ⇒ Sur-apprentissage du modèle
- ⇒ Plus de données (train) peut généraliser ou spécialiser
- ⇒ Attention aux effets de bord (données éparses)
- ⇒ Courbe **ROC** (précision / rappel)