Confidence Machines

- Introduction
- Principe Général
- Données Discrètes: exemple des SVM
- Données continues: exemple de la Ridge Regression
- Variante: Inductive Confidence Machine
- Conclusion
- Références

Confidence Machines: Introduction

- Objectifs recherchés:
  - Fournir un intervalle ou un niveau de confiance sur la réponse d'un modèle construit par apprentissage statistique.
  - Ne pas faire d'hypothèse « forte » sur les distributions des variables d'entrée et de sortie.
  - Résoudre le problème en un temps humainement acceptable.

Confidence Machines: Principe Général 1/4

- Problème posé:
  - On possède un ensemble d'entraînement $((\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_{m,}$ $y_m))$.
  - On cherche a estimer au mieux la réponse y_{m+1} correspondant à $\mathbf{x}_{m+1}$, et à fournir un niveau de confiance pour cette estimation.
- Hypothèse:
  - les $\mathbf{x}_i$ sont indépendamment et identiquement distribués

Confidence Machines: Principe Général 2/4

- Solution proposée
  - On résout le problème de manière transductive en intégrant le point $(\mathbf{x}_{m+1}, {}^*y_{m+1})$ à l'ensemble d'entraînement, ${}^*y_{m+1}$ étant une valeur *possible* de y_{m+1}.
  - Pour chaque valeur de ${}^*y_{m+1}$, on construit un nouveau modèle. Par exemple, dans le cas où $y_{m+1} \in \{A,B\}$, on construira un modèle entraîné sur l'ensemble $((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m),(\mathbf{x}_{m+1}, A))$ et un autre sur $((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m),(\mathbf{x}_{m+1},B))$
  - on calcul à quel point cette valeur est en **conformité** avec le reste de l'ensemble d'entraînement.
  - Pour chaque valeur de ${}^*y_{m+1}$, on a ainsi une mesure de la confiance que l'on a que ce soit la vraie valeur de y_{m+1}.
  - On peut en déduire un intervalle de confiance pour les valeurs continues, ou un niveau de confiance pour les valeurs discrètes.

Confidence Machines: Principe Général 3/4

- Conformité
  - On calcul la conformité d'un couple $z_{m+1} = (\mathbf{x}_{m+1}, {}^*y_{m+1})$ par la formule:

$$t(z_{1,\dots}z_{m+1}) = \frac{\#\{i=1,\dots,m+1 : \alpha_i \geq \alpha_{m+1}\}}{m+1}$$

    où # représente le nombre de fois que la proposition entre accolades est vérifiée. La conformité t est donc comprise entre 0 et 1.
  - On démontre que la probabilité que t soit inférieure à un seuil de confiance r est inférieure à r. Cela signifie que la valeur de t peut être utilisée comme mesure de probabilité maximale que $y_{m+1} = {}^*y_{m+1}$
  - Le coefficient α_i représente l'**étrangeté** de l'exemple z_i
- Etrangeté
  - Cette valeur mesure « l'écart » entre la réalité et la réponse du modèle.
  - Différentes mesures sont possibles en fonction de l'algorithme de modélisation utilisé, de la nature des données, etc.

Confidence Machines: Principe Général 4/4

- Avantages:
  - Pas d'hypothèse forte sur la distribution des données (seulement que les entrées soient i.i.d)
  - La mesure de conformité assure une certaine robustesse face à des modèles peu performants. Si les réponses sont toutes mauvaises de la même manière, l'ensemble reste cohérent !
  - Si les données sont effectivement i.i.d, la mesure de confiance est représentative de la distribution de la vraie valeur, et non de la valeur estimée par le modèle.
- Inconvénients:
  - La mesure d'étrangeté est cruciale et peut être difficile à calculer
  - L'approche transductive est coûteuse en terme de calcul
  - Cet algorithme paraît inapplicable pour prédire des intervalles sur des réponses continues: il faudrait réaliser une infinité de modèles !

Confidence Machines: Données Discrètes 1/2

- Exemple de mise en oeuvre en utilisant les SVM
  - $^*y_{m+1}$ peut prendre deux valeurs: {A; B}
  - les α_i correspondent aux poids que l'algorithme accorde aux données d'entraînement:
    - $\alpha_i = 0$ pour les exemples bien classés et loin de la marge
    - $\alpha_i > 0$ pour les vecteurs de support, c'est à dire les exemples situés à la marge (ou au delà dans le cas de l'algorithme « soft margin »)
  - En construisant deux modèles M_A et M_B pour notre exemple de test, on obtient deux valeurs de conformités: t_A et t_B.
  - On formule l'hypothèse que la réponse la plus plausible pour y_{m+1} est celle qui correspond à la conformité la plus élevée. Si par exemple on a $t_A > t_B$, on suppose que $y_{m+1} = A$

Confidence Machines: Données Discrètes 2/2

- Confiance en l'hypothèse
  - La **confiance** que l'on a sur l'hypothèse y_{m+1} = A est donnée par P{y_{m+1} = A | M_B}, soit la probabilité que y_{m+1} = A compte tenu des résultats du modèle pour lequel on avait fait l'hypothèse inverse.
  - Sachant que P{y_{m+1} = B | M_B} ≤ t_B, notre confiance est donc supérieure ou égale à (1 - t_B).
- Crédibilité de l'hypothèse
  - De même, on sait que P{y_{m+1} = A | M_A} ≤ t_A. Si cette valeur est faible (mais toujours supérieure à t_B), cela signifie que notre hypothèse y_{m+1} = A manque de **crédibilité**.
  - On va définir cette crédibilité comme valant simplement t_A
- Exemples:
  - tA = 0.9 et tB = 0.4: l'hypothèse y_{m+1} = A est sûre à 60% et crédible à 90%
  - tA = 0.4 et tB = 0.1: l'hypothèse y_{m+1} = A est sûre à 90% et crédible à 40%

Confidence Machines: Données Continues 1/2

- Exemple de mise en oeuvre en utilisant la Ridge Regression
  - La Ridge Regression est une méthode linéaire dans les paramètres pour laquelle on cherche le vecteur de poids **w** minimisant:

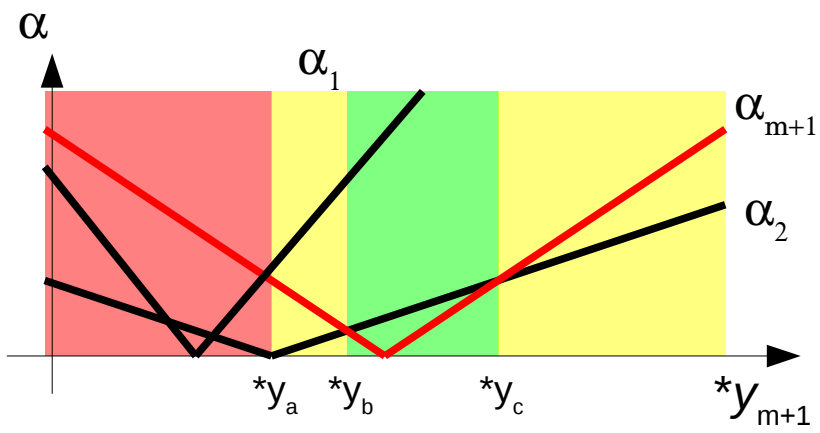
$$a\|\boldsymbol{w}\|^2+\|Y-X\boldsymbol{w}\|^2$$

  - La résolution de ce système nous donne: $\boldsymbol{w}=(X^T X+aI)^{-1}X^T Y$
  - Les coefficient d'étrangeté sont calculés ainsi: $\alpha_i=|y_i-\boldsymbol{w}.\boldsymbol{x_i}|$
  - Grâce à la linéarité de l'algorithme, il est possible d'exprimer le vecteur des α_i comme une expression linéaire de la variable $*y_{m+1}$:

$$\boldsymbol{\alpha}=|A+B\ *\ y_{m+1}|$$
$$avec\ A=(I-X(X^T X+aI)^{-1}X^T)(y_1,\ldots,y_m,0)^T$$
$$et\ B=(I-X(X^T X+aI)^{-1}X^T)(0,\ldots,0,1)^T$$

Confidence Machines: Données Continues 2/2

- Les valeurs des α_i varient linéairement par partie avec les valeurs de $^*y_{m+1}$, ce qui permet d'identifier un nombre fini de valeurs pour lesquelles la fonction de conformité t va changer.
- Il est alors possible de calculer analytiquement l'intervalle des $^*y_{m+1}$ dans lequel la vraie valeur y_{m+1} a au moins $r\%$ de chance de se trouver.
- Dans l'exemple ci-contre, on a:
  - $P\{y_{m+1} \leq {}^*y_a\} \leq 0.33$
  - $P\{y_{m+1} \in [{}^*y_a, {}^*y_b] \cup [{}^*y_c, \infty)\} \in [0.33, 0.67]$
  - $P\{y_{m+1} \in [{}^*y_b, {}^*y_c]\} \in [0.67, 1.0]$



Confidence Machines: Variante 1/2

- Afin de diminuer le nombre de calculs, une variante inductive de la méthode a été proposée pour les données continues.
  - On extrait un ensemble de « calibration » des m exemples d'entraînement, et on construit un modèle f à partir des k données d'entraînement restantes.
  - Les $\alpha_i = |y_i - f(\mathbf{x}_i)|$ sont calculés uniquement pour les m-(k+1) exemples de l'ensemble de calibration.
  - Les α_i étant calculés une fois pour toute, on peut leur associer une valeur de conformité:

$$t_i(z_{k+1}, \ldots z_m) = \frac{\#\{j = k+1, \ldots, m : \alpha_j \geq \alpha_i\}}{m-(k+1)}$$

  - Comme $\alpha_{m+1} = |y_{m+1} - f(\mathbf{x}_{m+1})|$, les valeurs de y_{m+1} permettant d'obtenir la conformité t_s, s ∈ {k+1, m}, se trouvent dans l'intervalle $[f(\mathbf{x}_{m+1}) - \alpha_s, f(\mathbf{x}_{m+1}) + \alpha_s]$.

Confidence Machines: Variante 2/2

- Avantages:
  - Le nombre de calculs nécessaires est considérablement réduit.
  - La méthode est indépendante de l'algorithme de modélisation sous-jacent.
- Inconvénients:
  - La méthode nécessite davantage de données d'entraînement.
  - Les α_i n'étant calculés que pour un seul modèle, la qualité prédictive de celui-ci influence directement la qualité des intervalles de confiance.
  - Pour un niveau de confiance donné, les intervalles sont à priori de taille égale pour tous les exemples de test, ce qui est peu crédible.
  - Les intervalles calculés semblent plus conservatifs (c'est à dire plus larges) que ceux de la méthode transductive.

Confidence Machines: Conclusion

- Les confidences machines sont basée sur une méthode simple et faisant très peu d'hypothèses sur la distribution des données.
- Elle est assez facile à mettre en oeuvre pour les réponses discrètes.
- C'est plus complexe pour les réponses continues:
  - difficultés pour calculer les coefficients d'étrangeté
  - dilemme nombre de calculs / qualité des intervalles
- Toutefois, dans ce second cas, des algorithmes permettent déjà de l'utiliser efficacement:
  - pour les ensembles de petite taille: Ridge Regression Confidence Machine (extensible aux données non linéaires grâce à la méthode des noyaux)
  - pour les ensembles de grande taille: Inductive Confidence Machine

Confidence Machines: Références

- Transduction with Confidence and Credibility (1999)
  - C. Saunders, A. Gammerman, V. Vovk
- Ridge Regression Confidence Machine (1999)
  - I. Nouretdinov, T. Melluish, V. Vovk
- Transductive Confidence Machine for Pattern Recognition (2001)
  - K. Proedrou, I. Nouretdinov, V. Vovk, A. Gammerman
- Inductive Confidence Machine for Regression (2001)
  - H. Papadopoulos, K. Proedrou, V. Vovk, A. Gammerman