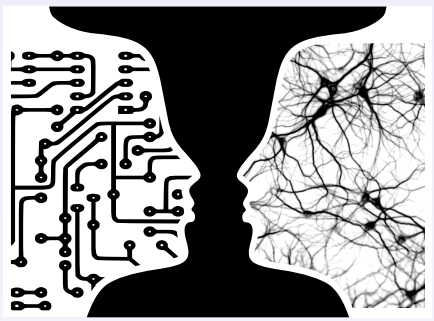
Algorithmes d'apprentissage et bornes sur le risque pour l'approche de régression à la prédiction de structures

Sébastien Giguère, François Laviolette,
Mario Marchand, and Khadidja Sylla



Groupe de Recherche en Apprentissage Automatique de Laval

Université Laval, Québec (QC)

septembre 2013

Plan

Plan

Espace de caractéristiques et noyaux

- Pour chaque exemple (x, y) : l'entrée x et la sortie y sont des objets structurés.
- **Caractéristiques d'entrée :** $\forall x \in \mathcal{X}: \quad x \mapsto X(x) \in \mathcal{H}_{\mathcal{X}}$ (RKHS)
- Produit scalaire : $\langle X(x)|X(x')\rangle = K_{\mathcal{X}}(x, x')$ (noyau d'entrée)

- **Caractéristiques de sortie :** $\forall y \in \mathcal{Y}: \quad y \mapsto Y(y) \in \mathcal{H}_{\mathcal{Y}}$ (RKHS)
- Produit scalaire : $\langle Y(y)|Y(y')\rangle = K_{\mathcal{Y}}(y, y')$ (noyau de sortie)

- Norme Euclidienne (au carré) $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle = K_{\mathcal{Y}}(y, y)$.

Espace de caractéristiques et noyaux

- Pour chaque exemple (x, y) : l'entrée x et la sortie y sont des objets structurés.
- **Caractéristiques d'entrée :** $\forall x \in \mathcal{X}: \quad x \mapsto X(x) \in \mathcal{H}_{\mathcal{X}}$ (RKHS)
- Produit scalaire : $\langle X(x)|X(x')\rangle = K_{\mathcal{X}}(x, x')$ (noyau d'entrée)

- **Caractéristiques de sortie :** $\forall y \in \mathcal{Y}: \quad y \mapsto Y(y) \in \mathcal{H}_{\mathcal{Y}}$ (RKHS)
- Produit scalaire : $\langle Y(y)|Y(y')\rangle = K_{\mathcal{Y}}(y, y')$ (noyau de sortie)

- Norme Euclidienne (au carré) $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle = K_{\mathcal{Y}}(y, y)$.

Espace de caractéristiques et noyaux

- Pour chaque exemple (x, y) : l'entrée x et la sortie y sont des objets structurés.
- **Caractéristiques d'entrée :** $\forall x \in \mathcal{X}: \quad x \mapsto X(x) \in \mathcal{H}_{\mathcal{X}}$ (RKHS)
- Produit scalaire : $\langle X(x)|X(x')\rangle = K_{\mathcal{X}}(x, x')$ (noyau d'entrée)

- **Caractéristiques de sortie :** $\forall y \in \mathcal{Y}: \quad y \mapsto Y(y) \in \mathcal{H}_{\mathcal{Y}}$ (RKHS)
- Produit scalaire : $\langle Y(y)|Y(y')\rangle = K_{\mathcal{Y}}(y, y')$ (noyau de sortie)

- Norme Euclidienne (au carré) $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle = K_{\mathcal{Y}}(y, y)$.

Espace de caractéristiques et noyaux

- Pour chaque exemple (x, y) : l'entrée x et la sortie y sont des objets structurés.
- **Caractéristiques d'entrée :** $\forall x \in \mathcal{X}: \quad x \mapsto X(x) \in \mathcal{H}_{\mathcal{X}}$ (RKHS)
- Produit scalaire : $\langle X(x)|X(x')\rangle = K_{\mathcal{X}}(x, x')$ (noyau d'entrée)

- **Caractéristiques de sortie :** $\forall y \in \mathcal{Y}: \quad y \mapsto Y(y) \in \mathcal{H}_{\mathcal{Y}}$ (RKHS)
- Produit scalaire : $\langle Y(y)|Y(y')\rangle = K_{\mathcal{Y}}(y, y')$ (noyau de sortie)

- Norme Euclidienne (au carré) $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle = K_{\mathcal{Y}}(y, y)$.

Espace de caractéristiques et noyaux

- Pour chaque exemple (x, y) : l'entrée x et la sortie y sont des objets structurés.
- **Caractéristiques d'entrée :** $\forall x \in \mathcal{X}: \quad x \mapsto X(x) \in \mathcal{H}_{\mathcal{X}}$ (RKHS)
- Produit scalaire : $\langle X(x)|X(x')\rangle = K_{\mathcal{X}}(x, x')$ (noyau d'entrée)

- **Caractéristiques de sortie :** $\forall y \in \mathcal{Y}: \quad y \mapsto Y(y) \in \mathcal{H}_{\mathcal{Y}}$ (RKHS)
- Produit scalaire : $\langle Y(y)|Y(y')\rangle = K_{\mathcal{Y}}(y, y')$ (noyau de sortie)

- Norme Euclidienne (au carré) $\|Y(y)\|^2 \stackrel{\text{def}}{=} \langle Y(y)|Y(y)\rangle = K_{\mathcal{Y}}(y, y)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Prédiction avec un opérateur linéaire

- Le **prédicteur W** est un **opérateur linéaire :** $\mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$.
- $\mathbf{W}X(x) =$ vecteur de caractéristiques de sortie **prédit** pour l'entrée x.
- La sortie prédite $y_{\mathbf{W}}(x)$ pour l'entrée x est donnée par

$$y_{\mathbf{W}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathcal{Y}} \|Y(y) - \mathbf{W}X(x)\|^2 .$$

- $y_{\mathbf{W}}(x)$ nécessite uniquement $K_{\mathcal{X}}$ et $K_{\mathcal{Y}}$ (à la place de X et Y) lorsque

$$\mathbf{W}X(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) A_{i,j} K_{\mathcal{X}}(x_j, x) .$$

- Trouver $y_{\mathbf{W}}(x)$ est un **problème de pré-image** souvent $\mathcal{NP}$-difficile.
- Utilisons une fonction de perte de régression qui ne dépend pas $y_{\mathbf{W}}(x)$.

Fonction de perte du noyau de sortie et perte quadratique

$K_{\mathcal{Y}}$ induit une **fonction de perte du noyau de sortie** $L_{K_{\mathcal{Y}}}$ définie par

$$\begin{aligned} L_{K_{\mathcal{Y}}}(y, y') &\stackrel{\text{def}}{=} \frac{1}{2}\|Y(y) - Y(y')\|^2 \\ &= \frac{1}{2}\left[K_{\mathcal{Y}}(y, y) + K_{\mathcal{Y}}(y', y')\right] - K_{\mathcal{Y}}(y, y')\,. \end{aligned}$$

Lemme

Pour tout prédicteur $\mathbf{W}$, pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, nous avons

$$L_{K_{\mathcal{Y}}}(y_{\mathbf{W}}(x), y) \;\leq\; 2\,\|Y(y) - \mathbf{W}X(x)\|^2\;.$$

Notez que $\|Y(y) - \mathbf{W}X(x)\|^2$ **ne dépend pas de** $y_{\mathbf{W}}(x)$.

Fonction de perte du noyau de sortie et perte quadratique

$K_{\mathcal{Y}}$ induit une **fonction de perte du noyau de sortie** $L_{K_{\mathcal{Y}}}$ définie par

$$\begin{aligned} L_{K_{\mathcal{Y}}}(y, y') &\stackrel{\text{def}}{=} \frac{1}{2}\|Y(y) - Y(y')\|^2 \\ &= \frac{1}{2}\left[K_{\mathcal{Y}}(y, y) + K_{\mathcal{Y}}(y', y')\right] - K_{\mathcal{Y}}(y, y')\,. \end{aligned}$$

Lemme

Pour tout prédicteur $\mathbf{W}$, pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, nous avons

$$L_{K_{\mathcal{Y}}}(y_{\mathbf{W}}(x), y) \;\leq\; 2\,\|Y(y) - \mathbf{W}X(x)\|^2\,.$$

Notez que $\|Y(y) - \mathbf{W}X(x)\|^2$ **ne dépend pas de** $y_{\mathbf{W}}(x)$.

Perte de prédiction et perte quadratique

Démonstration.

De l'inégalité triangulaire, nous avons

$$\|Y(y) - Y(y_{\mathbf{w}}(x))\| \leq \|Y(y) - \mathbf{W}X(x)\| + \|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\|$$

Nous avons aussi $\|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\| \leq \|Y(y) - \mathbf{W}X(x)\|$. □

Si $K_{\mathcal{Y}}$ est tel que la **perte de prédiction** $L(y, y') \leq L_{K_Y}(y, y')$, alors

$$L(y_{\mathbf{w}}(x), y) \leq L_{K_{\mathcal{Y}}}(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2 .$$

Alors

$$\mathop{\mathbf{E}}_{(x,y)\sim D} \|Y(y) - \mathbf{W}X(x)\|^2 \text{ petit} \implies \mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \text{ petit} .$$

Nous pouvons alors minimiser risque quadratique à la place du risque de prédiction.

Perte de prédiction et perte quadratique

Démonstration.

De l'inégalité triangulaire, nous avons

$$\|Y(y) - Y(y_{\mathbf{w}}(x))\| \leq \|Y(y) - \mathbf{W}X(x)\| + \|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\|$$

Nous avons aussi $\|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\| \leq \|Y(y) - \mathbf{W}X(x)\|$. □

Si $K_{\mathcal{Y}}$ est tel que la **perte de prédiction** $L(y, y') \leq L_{K_Y}(y, y')$, alors

$$L(y_{\mathbf{w}}(x), y) \leq L_{K_{\mathcal{Y}}}(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2 .$$

Alors

$$\mathop{\mathbf{E}}_{(x,y)\sim D} \|Y(y) - \mathbf{W}X(x)\|^2 \text{ petit} \implies \mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \text{ petit}.$$

Nous pouvons alors minimiser risque quadratique à la place du risque de prédiction.

Perte de prédiction et perte quadratique

Démonstration.

De l'inégalité triangulaire, nous avons

$$\|Y(y) - Y(y_{\mathbf{w}}(x))\| \leq \|Y(y) - \mathbf{W}X(x)\| + \|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\|$$

Nous avons aussi $\|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\| \leq \|Y(y) - \mathbf{W}X(x)\|$. □

Si $K_{\mathcal{Y}}$ est tel que la **perte de prédiction** $L(y, y') \leq L_{K_Y}(y, y')$, alors

$$L(y_{\mathbf{w}}(x), y) \leq L_{K_{\mathcal{Y}}}(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2 .$$

Alors

$$\mathop{\mathbf{E}}_{(x,y)\sim D} \|Y(y) - \mathbf{W}X(x)\|^2 \text{ petit} \implies \mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \text{ petit} .$$

Nous pouvons alors minimiser risque quadratique à la place du risque de prédiction.

Perte de prédiction et perte quadratique

Démonstration.

De l'inégalité triangulaire, nous avons

$$\|Y(y) - Y(y_{\mathbf{w}}(x))\| \leq \|Y(y) - \mathbf{W}X(x)\| + \|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\|$$

Nous avons aussi $\|Y(y_{\mathbf{w}}(x)) - \mathbf{W}X(x)\| \leq \|Y(y) - \mathbf{W}X(x)\|$. □

Si $K_{\mathcal{Y}}$ est tel que la **perte de prédiction** $L(y, y') \leq L_{K_Y}(y, y')$, alors

$$L(y_{\mathbf{w}}(x), y) \leq L_{K_{\mathcal{Y}}}(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2 .$$

Alors

$$\mathop{\mathbf{E}}_{(x,y)\sim D} \|Y(y) - \mathbf{W}X(x)\|^2 \text{ petit} \implies \mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \text{ petit} .$$

Nous pouvons alors minimiser risque quadratique à la place du risque de prédiction.

Minimisation du risque quadratique

Minimiser $\underset{(x,y)\sim D}{\mathbf{E}} \|Y(y) - \mathbf{W}X(x)\|^2$ à la place de $\underset{(x,y)\sim D}{\mathbf{E}} L(y_{\mathbf{w}}(x), y)$.

AVANTAGE :

- Pas nécessaire de calculer $y_{\mathbf{w}}(x)$ pour chaque x de l'échantillon d'apprentissage et pour chaque mise à jour de $\mathbf{W}$: un algorithme d'apprentissage beaucoup plus rapide.

DÉSAVANTAGES :

- Inconsistance : il existe D tel que le minimiseur du risque quadratique possède un risque de prédiction plus élevé que le minimiseur du risque de prédiction (résultat connu en classification binaire).
- Pour certains risques de prédiction L, il peut s'avérer difficile de trouver $K_{\mathcal{Y}}$ tel que $L(y, y') \leq L_{K_{\mathcal{Y}}}(y, y')$ (avec un majorant précis).

Minimisation du risque quadratique

Minimiser $\underset{(x,y)\sim D}{\mathbf{E}} \|Y(y) - \mathbf{W}X(x)\|^2$ à la place de $\underset{(x,y)\sim D}{\mathbf{E}} L(y_{\mathbf{w}}(x), y)$.

AVANTAGE :

- Pas nécessaire de calculer $y_{\mathbf{w}}(x)$ pour chaque x de l'échantillon d'apprentissage et pour chaque mise à jour de $\mathbf{W}$: un algorithme d'apprentissage beaucoup plus rapide.

DÉSAVANTAGES :

- Inconsistance : il existe D tel que le minimiseur du risque quadratique possède un risque de prédiction plus élevé que le minimiseur du risque de prédiction (résultat connu en classification binaire).
- Pour certains risques de prédiction L, il peut s'avérer difficile de trouver $K_{\mathcal{Y}}$ tel que $L(y, y') \leq L_{K_{\mathcal{Y}}}(y, y')$ (avec un majorant précis).

Minimisation du risque quadratique

Minimiser $\underset{(x,y)\sim D}{\mathbf{E}} \|Y(y) - \mathbf{W}X(x)\|^2$ à la place de $\underset{(x,y)\sim D}{\mathbf{E}} L(y_{\mathbf{w}}(x), y)$.

AVANTAGE :

- Pas nécessaire de calculer $y_{\mathbf{w}}(x)$ pour chaque x de l'échantillon d'apprentissage et pour chaque mise à jour de $\mathbf{W}$: un algorithme d'apprentissage beaucoup plus rapide.

DÉSAVANTAGES :

- Inconsistance : il existe D tel que le minimiseur du risque quadratique possède un risque de prédiction plus élevé que le minimiseur du risque de prédiction (résultat connu en classification binaire).
- Pour certains risques de prédiction L, il peut s'avérer difficile de trouver $K_{\mathcal{Y}}$ tel que $L(y, y') \leq L_{K_{\mathcal{Y}}}(y, y')$ (avec un majorant précis).

Minimisation du risque quadratique

Minimiser $\underset{(x,y)\sim D}{\mathbf{E}} \|Y(y) - \mathbf{W}X(x)\|^2$ à la place de $\underset{(x,y)\sim D}{\mathbf{E}} L(y_{\mathbf{w}}(x), y)$.

AVANTAGE :

- Pas nécessaire de calculer $y_{\mathbf{w}}(x)$ pour chaque x de l'échantillon d'apprentissage et pour chaque mise à jour de $\mathbf{W}$: un algorithme d'apprentissage beaucoup plus rapide.

DÉSAVANTAGES :

- Inconsistance : il existe D tel que le minimiseur du risque quadratique possède un risque de prédiction plus élevé que le minimiseur du risque de prédiction (résultat connu en classification binaire).
- Pour certains risques de prédiction L, il peut s'avérer difficile de trouver $K_{\mathcal{Y}}$ tel que $L(y, y') \leq L_{K_{\mathcal{Y}}}(y, y')$ (avec un majorant précis).

Minimisation du risque quadratique

Minimiser $\underset{(x,y)\sim D}{\mathbf{E}} \|Y(y) - \mathbf{W}X(x)\|^2$ à la place de $\underset{(x,y)\sim D}{\mathbf{E}} L(y_{\mathbf{w}}(x), y)$.

AVANTAGE :

- Pas nécessaire de calculer $y_{\mathbf{w}}(x)$ pour chaque x de l'échantillon d'apprentissage et pour chaque mise à jour de $\mathbf{W}$: un algorithme d'apprentissage beaucoup plus rapide.

DÉSAVANTAGES :

- Inconsistance : il existe D tel que le minimiseur du risque quadratique possède un risque de prédiction plus élevé que le minimiseur du risque de prédiction (résultat connu en classification binaire).
- Pour certains risques de prédiction L, il peut s'avérer difficile de trouver $K_{\mathcal{Y}}$ tel que $L(y, y') \leq L_{K_{\mathcal{Y}}}(y, y')$ (avec un majorant précis).

Plan

Une première borne sur le risque

Si $L_{K_\mathcal{Y}}$ majore L, on a $L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$. Donc $\forall a \geq 1$,

$$L(y_{\mathbf{w}}(x), y) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{1}{a}L(y_{\mathbf{w}}(x),y)}\right) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{2}{a}\|Y(y)-\mathbf{W}X(x)\|^2}\right)$$

En majorant l'espérance sur (x, y) du terme de droite, on obtient

Theorem

Soit $K_\mathcal{X}(x, x) = 1\ \forall x \in \mathcal{X}$. Soit $\mathcal{H}_\mathcal{X}$ et $\mathcal{H}_\mathcal{Y}$ de dimension finie. Avec probabilité $\geq 1 - \delta$ sur tous les échantillons $S \sim D^m$, nous avons simultanément pour tout $\mathbf{W}$,

$$\mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \leq \frac{5e}{e-1}\left[1 - e^{-\frac{1}{m}\left(2\sum_{i=1}^m \|Y(y_i)-\mathbf{W}X(x_i)\|^2 + \frac{9}{8}\|\mathbf{W}\|^2 + \ln\frac{1}{\delta}\right)}\right].$$

Une première borne sur le risque

Si $L_{K_\mathcal{Y}}$ majore L, on a $L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$. Donc $\forall a \geq 1$,

$$L(y_{\mathbf{w}}(x), y) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{1}{a}L(y_{\mathbf{w}}(x),y)}\right) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{2}{a}\|Y(y)-\mathbf{W}X(x)\|^2}\right)$$

En majorant l'espérance sur (x, y) du terme de droite, on obtient

Theorem

Soit $K_\mathcal{X}(x, x) = 1\ \forall x \in \mathcal{X}$. Soit $\mathcal{H}_\mathcal{X}$ et $\mathcal{H}_\mathcal{Y}$ de dimension finie. Avec probabilité $\geq 1 - \delta$ sur tous les échantillons $S \sim D^m$, nous avons simultanément pour tout $\mathbf{W}$,

$$\mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \leq \frac{5e}{e-1}\left[1 - e^{-\frac{1}{m}\left(2\sum_{i=1}^m \|Y(y_i)-\mathbf{W}X(x_i)\|^2 + \frac{9}{8}\|\mathbf{W}\|^2 + \ln\frac{1}{\delta}\right)}\right].$$

Une première borne sur le risque

Si $L_{K_{\mathcal{Y}}}$ majore L, on a $L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$. Donc $\forall a \geq 1$,

$$L(y_{\mathbf{w}}(x), y) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{1}{a}L(y_{\mathbf{w}}(x),y)}\right) \leq \frac{ae}{e-1}\left(1 - e^{-\frac{2}{a}\|Y(y)-\mathbf{W}X(x)\|^2}\right)$$

En majorant l'espérance sur (x, y) du terme de droite, on obtient

Theorem

Soit $K_{\mathcal{X}}(x, x) = 1\ \forall x \in \mathcal{X}$. Soit $\mathcal{H}_{\mathcal{X}}$ et $\mathcal{H}_{\mathcal{Y}}$ de dimension finie. Avec probabilité $\geq 1 - \delta$ sur tous les échantillons $S \sim D^m$, nous avons simultanément pour tout $\mathbf{W}$,

$$\mathop{\mathbf{E}}_{(x,y)\sim D} L(y_{\mathbf{w}}(x), y) \leq \frac{5e}{e-1}\left[1 - e^{-\frac{1}{m}\left(2\sum_{i=1}^{m}\|Y(y_i)-\mathbf{W}X(x_i)\|^2 + \frac{9}{8}\|\mathbf{W}\|^2 + \ln\frac{1}{\delta}\right)}\right].$$

Le minimiseur de cette borne

Le prédicteur minimisant cette borne est celui minimisant la fonctionnelle proposée par Cortes, Mohri, et Weston (2007) :

$$C \sum_{i=1}^{m} \| Y(y_i) - \mathbf{W} X(x_i) \|^2 + \| \mathbf{W} \|^2 ,$$

pour $C > 0$. Le minimiseur $\mathbf{W}^*$ est unique (si $C < \infty$) et donné par

$$\mathbf{W}^* = \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) \left(\mathbf{K}_{\mathcal{X}} + \frac{1}{C} \mathbf{I} \right)^{-1}_{i,j} X^\top(x_j) ,$$

où $X^\top(x)$ denote la transpose de $X(x)$, $\mathbf{K}_{\mathcal{X}}$ denote la matrice du noyau d'entrée et $\mathbf{I}$ denote la matrice identité $m \times m$.

Note : $\mathbf{W}^*$ est une combinaison linéaire d'operateurs $Y(y_i)X^\top(x_j)$

Le minimiseur de cette borne

Le prédicteur minimisant cette borne est celui minimisant la fonctionnelle proposée par Cortes, Mohri, et Weston (2007) :

$$C \sum_{i=1}^{m} \| Y(y_i) - \mathbf{W} X(x_i) \|^2 + \|\mathbf{W}\|^2 \,,$$

pour $C > 0$. Le minimiseur $\mathbf{W}^*$ est unique (si $C < \infty$) et donné par

$$\mathbf{W}^* \;=\; \sum_{i=1}^{m} \sum_{j=1}^{m} Y(y_i) \left(\mathbf{K}_{\mathcal{X}} + \frac{1}{C} \mathbf{I} \right)^{-1}_{i,j} X^\top(x_j) \,,$$

où $X^\top(x)$ denote la transpose de $X(x)$, $\mathbf{K}_{\mathcal{X}}$ denote la matrice du noyau d'entrée et $\mathbf{I}$ denote la matrice identité $m \times m$.

Note : $\mathbf{W}^*$ est une combinaison linéaire d'operateurs $Y(y_i)X^\top(x_j)$

Plan

Combinaisons linéaires d'opérateurs à deux exemples

Considérons une combinaison linéaire arbitraire des opérateurs $Y(y_i)X^\intercal(x_j)$:

$$\mathbf{W} = \sum_{i=1}^{m}\sum_{j=1}^{m} Y(y_i)A_{i,j}X^\intercal(x_j)\,,$$

La perte quadratique $\|Y(y) - \mathbf{W}X(x)\|^2$ est maintenant donnée par

$$\left\| Y(y) - \sum_{i=1}^{m}\sum_{j=1}^{m} A_{i,j}K_{\mathcal{X}}(x_j, x)Y(y_i) \right\|^2 \stackrel{\text{def}}{=} R(\mathbf{A}, x, y)\,.$$

Soit

$$R(\mathbf{A}) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} R(\mathbf{A}, x, y) \quad ; \quad R(\mathbf{A}, S) = \frac{1}{m}\sum_{i=1}^{m} R(\mathbf{A}, x_i, y_i)$$

Combinaisons linéaires d'opérateurs à deux exemples

Considérons une combinaison linéaire arbitraire des opérateurs $Y(y_i)X^\intercal(x_j)$:

$$\mathbf{W} = \sum_{i=1}^{m}\sum_{j=1}^{m} Y(y_i)A_{i,j}X^\intercal(x_j)\,,$$

La perte quadratique $\|Y(y) - \mathbf{W}X(x)\|^2$ est maintenant donnée par

$$\left\| Y(y) - \sum_{i=1}^{m}\sum_{j=1}^{m} A_{i,j}K_{\mathcal{X}}(x_j, x)Y(y_i) \right\|^2 \stackrel{\text{def}}{=} R(\mathbf{A}, x, y)\,.$$

Soit

$$R(\mathbf{A}) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} R(\mathbf{A}, x, y) \quad ; \quad R(\mathbf{A}, S) = \frac{1}{m}\sum_{i=1}^{m} R(\mathbf{A}, x_i, y_i)$$

Combinaisons linéaires d'opérateurs à deux exemples

Considérons une combinaison linéaire arbitraire des opérateurs $Y(y_i)X^\intercal(x_j)$:

$$\mathbf{W} = \sum_{i=1}^{m}\sum_{j=1}^{m} Y(y_i)A_{i,j}X^\intercal(x_j)\,,$$

La perte quadratique $\|Y(y) - \mathbf{W}X(x)\|^2$ est maintenant donnée par

$$\left\| Y(y) - \sum_{i=1}^{m}\sum_{j=1}^{m} A_{i,j}K_{\mathcal{X}}(x_j, x)Y(y_i) \right\|^2 \stackrel{\text{def}}{=} R(\mathbf{A}, x, y)\,.$$

Soit

$$R(\mathbf{A}) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} R(\mathbf{A}, x, y) \quad ; \quad R(\mathbf{A}, S) = \frac{1}{m}\sum_{i=1}^{m} R(\mathbf{A}, x_i, y_i)$$

Une borne PAC-Bayes avec compression d'échantillon

Theorem

Avec probabilité $\geq 1-\delta$ *sur tous les échantillons* $S \sim D^m$, *on a simultanément pour tout* $\mathbf{A}$

$$R(\mathbf{A}) \leq R(\mathbf{A}, S) + \sqrt{\frac{2B_{\mathcal{Y}}(1+\kappa B_{\mathcal{X}})^2}{2(m-4)}\left[20 + \ln\left(\frac{8\sqrt{m}}{\delta}\right)\right]}\,.$$

avec $|K_{\mathcal{Y}}(y,y')| \leq B_{\mathcal{Y}}$ *et* $|K_{\mathcal{X}}(x,x')| \leq B_{\mathcal{X}}$ *et*

$$A_{i,j} = \kappa(q^+_{i,j} - q^-_{i,j}) \quad \textit{avec} \sum_{i=1}^{m}\sum_{j=1}^{m}(q^+_{i,j} + q^-_{i,j}) = 1\,.$$

Valide même lorsque $\mathcal{H}_{\mathcal{X}}$ sont $\mathcal{H}_{\mathcal{Y}}$ sont des RKHS de dimension infini.

Idée de la preuve

$q^{\pm}_{i,j}$ est le poids assigné au predicteur $\pm Y(y_i)X^{\intercal}(x_j)$. Alors

$$R(\mathbf{A}, x, y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} q^s_{\mathbf{i}}\, q^t_{\mathbf{j}}\, \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x, y)\,,$$

où $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, m\}^2$ est l'ensemble des paires d'indices et $\mathcal{W} \stackrel{\text{def}}{=} \{-1, +1\}$.
Puisqu'il s'agit d'un risque de Gibbs, nous pouvons utiliser PAC-Bayes.
Le prior $\mathbf{p}$ est uniforme sur $\mathcal{I} \times \mathcal{W}$ et $\mathbf{q}$ est **quasi uniforme** sur $\mathcal{I} \times \mathcal{W}$:

$$q^{+}_{i,j} + q^{-}_{i,j} = \frac{1}{m^2} \quad \Longrightarrow \quad \mathrm{KL}(\mathbf{q}, \mathbf{p}) \leq \ln(2)\,.$$

Idée de la preuve

$q^{\pm}_{i,j}$ est le poids assigné au predicteur $\pm Y(y_i)X^{\intercal}(x_j)$. Alors

$$R(\mathbf{A}, x, y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{I}} \sum_{s\in\mathcal{W}} \sum_{t\in\mathcal{W}} q^{s}_{\mathbf{i}}\, q^{t}_{\mathbf{j}}\, \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x, y)\,,$$

où $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, m\}^2$ est l'ensemble des paires d'indices et $\mathcal{W} \stackrel{\text{def}}{=} \{-1, +1\}$.
Puisqu'il s'agit d'un risque de Gibbs, nous pouvons utiliser PAC-Bayes.
Le prior $\mathbf{p}$ est uniforme sur $\mathcal{I} \times \mathcal{W}$ et $\mathbf{q}$ est **quasi uniforme** sur $\mathcal{I} \times \mathcal{W}$:

$$q^{+}_{i,j} + q^{-}_{i,j} = \frac{1}{m^2} \quad \Longrightarrow \quad \mathrm{KL}(\mathbf{q}, \mathbf{p}) \leq \ln(2)\,.$$

Idée de la preuve

$q^{\pm}_{i,j}$ est le poids assigné au predicteur $\pm Y(y_i)X^{\intercal}(x_j)$. Alors

$$R(\mathbf{A}, x, y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{I}} \sum_{s\in\mathcal{W}} \sum_{t\in\mathcal{W}} q^{s}_{\mathbf{i}}\, q^{t}_{\mathbf{j}}\, \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x, y)\,,$$

où $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, m\}^2$ est l'ensemble des paires d'indices et $\mathcal{W} \stackrel{\text{def}}{=} \{-1, +1\}$.

Puisqu'il s'agit d'un risque de Gibbs, nous pouvons utiliser PAC-Bayes.

Le prior $\mathbf{p}$ est uniforme sur $\mathcal{I} \times \mathcal{W}$ et $\mathbf{q}$ est **quasi uniforme** sur $\mathcal{I} \times \mathcal{W}$:

$$q^{+}_{i,j} + q^{-}_{i,j} = \frac{1}{m^2} \quad \Longrightarrow \quad \mathrm{KL}(\mathbf{q}, \mathbf{p}) \leq \ln(2)\,.$$

Idée de la preuve

$q^{\pm}_{i,j}$ est le poids assigné au predicteur $\pm Y(y_i)X^{\intercal}(x_j)$. Alors

$$R(\mathbf{A}, x, y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{I}} \sum_{s\in\mathcal{W}} \sum_{t\in\mathcal{W}} q^{s}_{\mathbf{i}}\, q^{t}_{\mathbf{j}}\, \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x, y)\,,$$

où $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, m\}^2$ est l'ensemble des paires d'indices et $\mathcal{W} \stackrel{\text{def}}{=} \{-1, +1\}$. Puisqu'il s'agit d'un risque de Gibbs, nous pouvons utiliser PAC-Bayes.

Le prior $\mathbf{p}$ est uniforme sur $\mathcal{I} \times \mathcal{W}$ et $\mathbf{q}$ est **quasi uniforme** sur $\mathcal{I} \times \mathcal{W}$:

$$q^{+}_{i,j} + q^{-}_{i,j} = \frac{1}{m^2} \quad \Longrightarrow \quad \mathrm{KL}(\mathbf{q}, \mathbf{p}) \leq \ln(2)\,.$$

Idée de la preuve

$q^{\pm}_{i,j}$ est le poids assigné au predicteur $\pm Y(y_i)X^{\intercal}(x_j)$. Alors

$$R(\mathbf{A}, x, y) \;=\; \sum_{\mathbf{i}\in\mathcal{I}}\sum_{\mathbf{j}\in\mathcal{I}}\sum_{s\in\mathcal{W}}\sum_{t\in\mathcal{W}} q^{s}_{\mathbf{i}}\, q^{t}_{\mathbf{j}}\, \ell^{s,t}_{\mathbf{i},\mathbf{j}}(x, y)\,,$$

où $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, m\}^2$ est l'ensemble des paires d'indices et $\mathcal{W} \stackrel{\text{def}}{=} \{-1, +1\}$.
Puisqu'il s'agit d'un risque de Gibbs, nous pouvons utiliser PAC-Bayes.
Le prior $\mathbf{p}$ est uniforme sur $\mathcal{I} \times \mathcal{W}$ et $\mathbf{q}$ est **quasi uniforme** sur $\mathcal{I} \times \mathcal{W}$:

$$q^{+}_{i,j} + q^{-}_{i,j} = \frac{1}{m^2} \quad \Longrightarrow \quad \mathrm{KL}(\mathbf{q}, \mathbf{p}) \leq \ln(2)\,.$$

Le problème d'optimisation

Puisque $A_{i,j} = \kappa(q^+_{i,j} - q^-_{i,j})$, et que $\mathbf{q}$ est quasi uniforme, on a

$$|A_{i,j}| \leq C \ \forall (i,j) \quad \text{pour un } C > 0\,.$$

Computationellement avantageux de replacer ces m^2 contraintes, par la seule contrainte

$$\sum_{(i,j)\in\mathcal{I}} A^2_{i,j} \leq R^2 \quad \text{pour un } R > 0\,.$$

Notez que $|A_{i,j}| \leq R$ pour tout (i,j) lorsque cette contrainte ℓ_2 est satisfaite. Alors, pour $R > 0$, résolvons

$$\min_{\mathbf{A}} R(\mathbf{A}, S) \quad \text{s.t. } \sum_{i=1}^{m}\sum_{j=1}^{m} A^2_{i,j} \leq R^2\,.$$

Le problème d'optimisation

Puisque $A_{i,j} = \kappa(q^+_{i,j} - q^-_{i,j})$, et que $\mathbf{q}$ est quasi uniforme, on a

$$|A_{i,j}| \leq C \ \forall (i,j) \quad \text{pour un } C > 0\,.$$

Computationellement avantageux de replacer ces m^2 contraintes, par la seule contrainte

$$\sum_{(i,j)\in\mathcal{I}} A^2_{i,j} \leq R^2 \quad \text{pour un } R > 0\,.$$

Notez que $|A_{i,j}| \leq R$ pour tout (i,j) lorsque cette contrainte ℓ_2 est satisfaite. Alors, pour $R > 0$, résolvons

$$\min_{\mathbf{A}} R(\mathbf{A}, S) \quad \text{s.t.} \sum_{i=1}^{m}\sum_{j=1}^{m} A^2_{i,j} \leq R^2\,.$$

Le problème d'optimisation

Puisque $A_{i,j} = \kappa(q^+_{i,j} - q^-_{i,j})$, et que $\mathbf{q}$ est quasi uniforme, on a

$$|A_{i,j}| \leq C \ \forall (i,j) \quad \text{pour un } C > 0\,.$$

Computationellement avantageux de replacer ces m^2 contraintes, par la seule contrainte

$$\sum_{(i,j)\in\mathcal{I}} A^2_{i,j} \leq R^2 \quad \text{pour un } R > 0\,.$$

Notez que $|A_{i,j}| \leq R$ pour tout (i,j) lorsque cette contrainte ℓ_2 est satisfaite. Alors, pour $R > 0$, résolvons

$$\min_{\mathbf{A}} R(\mathbf{A}, S) \quad \text{s.t.} \ \sum_{i=1}^{m}\sum_{j=1}^{m} A^2_{i,j} \leq R^2\,.$$

La solution (obtenue en temps $\in O(m^3)$)

Theorem

Soit l'ensemble $\mathcal{A}^*$ des solutions. Soit $v_1, \ldots, v_m$ et $\lambda_1, \ldots, \lambda_m$ les vecteurs propres et valeurs propres de $\mathbf{K}_{\mathcal{X}}$. Soit $u_1, \ldots, u_m$ et $\delta_1, \ldots, \delta_m$ les vecteurs propres et valeurs propres de $\mathbf{K}_{\mathcal{Y}}$. Soit $\mathcal{J} \stackrel{def}{=} \{(i,j) \in \mathcal{I} : \delta_i \lambda_j > 0\}$. Alors $\sum_{i=1}^m \sum_{j=1}^m \gamma_{i,j} u_i v_j^\top \in \mathcal{A}^*$, où $\gamma_{i,j}$ est donné par

$$\text{Si} \sum_{(i,j) \in \mathcal{J}} \frac{(u_i^\top v_j)^2}{\lambda_j^2} \leq R^2 \text{ alors } \gamma_{i,j} = \begin{cases} 0 & \text{si } \delta_i \lambda_j = 0 \\ \frac{u_i^\top v_j}{\lambda_j} & \text{si } \delta_i \lambda_j > 0 \end{cases}$$

$$\text{Autrement } \gamma_{i,j} = \frac{\delta_i \lambda_j (u_i^\top v_j)}{\delta_i \lambda_j^2 + m\beta} \,,$$

où $\beta > 0$ est solution de $\sum_{i=1}^m \sum_{j=1}^m \frac{\delta_i^2 \lambda_j^2 (u_i^\top v_j)^2}{(\delta_i \lambda_j^2 + m\beta)^2} = R^2$.

Plan

Deux algorithmes d'apprentissage testés sur deux tâches

- Le minimiseur de la 1ière borne : SORR (ridge regression)
- Le minimiseur de la 2e borne : SOSC (sample-compression)
- Tâche de reconnaissance de mots manuscrits : le protocole utilisé est celui de Taskar et al. (2004).
- Tâche de classification hiérarchique d'enzymes : le protocole utilisé est celui de Rousu et al. (2006).

Deux algorithmes d'apprentissage testés sur deux tâches

- Le minimiseur de la 1ière borne : SORR (ridge regression)
- Le minimiseur de la 2e borne : SOSC (sample-compression)
- Tâche de reconnaissance de mots manuscrits : le protocole utilisé est celui de Taskar et al. (2004).
- Tâche de classification hiérarchique d'enzymes : le protocole utilisé est celui de Rousu et al. (2006).

Deux algorithmes d'apprentissage testés sur deux tâches

- Le minimiseur de la 1ière borne : SORR (ridge regression)
- Le minimiseur de la 2e borne : SOSC (sample-compression)
- Tâche de reconnaissance de mots manuscrits : le protocole utilisé est celui de Taskar et al. (2004).
- Tâche de classification hiérarchique d'enzymes : le protocole utilisé est celui de Rousu et al. (2006).

Deux algorithmes d'apprentissage testés sur deux tâches

- Le minimiseur de la 1ière borne : SORR (ridge regression)
- Le minimiseur de la 2e borne : SOSC (sample-compression)
- Tâche de reconnaissance de mots manuscrits : le protocole utilisé est celui de Taskar et al. (2004).
- Tâche de classification hiérarchique d'enzymes : le protocole utilisé est celui de Rousu et al. (2006).

Reconnaissance de mots manuscrits

- Entrée : image du mot manuscrit.
- Sortie : séquence de caractères du mot.
- Noyau d'entrée : polynomial de degré d
- Noyaux de sortie : Dirac et Hamming.
- Le noyau de Dirac est le meilleur sur la perte 0/1.
- Le noyau de Hamming devrait être le meilleur pour la perte "letter".

	Dirac kernel		Hamming kernel	
	SORR	SOSC	SORR	SOSC
0/1 risk	0.0539 $\pm$.0087	0.0525 $\pm$.0085	0.0871 $\pm$.0078	0.0871 $\pm$.0078
Letter risk	0.0294 $\pm$.0067	0.0285 $\pm$.0062	0.0370 $\pm$.0047	0.0367 $\pm$.0049

Classification hiérarchique d'enzymes

- Entrée : la séquence d'acides aminées de l'enzyme.
- Sortie : le chemin dans l'arbre de classification des enzymes.
- Noyau d'entrée = 4-gram (pour tous les algorithmes)
- Noyau de sortie = noyau hiérarchique (longueur du sous-chemin commun)
- H-risk : longueur de la partie du chemin en désaccord

	H–M^3–ℓ_Δ	H–M^3–$\ell_{\tilde{H}}$	SORR	SOSC
0/1	0.957 [0.949, 0.965]	0.855 [0.840, 0.869]	0.640 [0.621, 0.659]	0.684 [0.666, 0.702]
H risk	1.2	2.50	1.71	1.84

Plan

Conclusion

- L majoré par $L_{\mathcal{K}_\mathcal{Y}} \Rightarrow L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$.
- Dans ce cas, l'approche par régression est justifiée : deux bornes sur le risque sont proposées.
- Le minimiseur de la 1ière borne, SORR, est l'estimateur des moindres carrés (régularisé) étudié par Cortes, Mohri, et Weston (2007).
- Le minimiseur de la 2e borne, SOSC, est nouveau.
- L'approche par régression n'est présentement pas justifiée lorsque L n'est pas majoré par $L_{K_\mathcal{Y}}$.

Conclusion

- L majoré par $L_{\mathcal{K}_\mathcal{Y}} \Rightarrow L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$.
- Dans ce cas, l'approche par régression est justifiée : deux bornes sur le risque sont proposées.
- Le minimiseur de la 1ière borne, SORR, est l'estimateur des moindres carrés (régularisé) étudié par Cortes, Mohri, et Weston (2007).
- Le minimiseur de la 2e borne, SOSC, est nouveau.
- L'approche par régression n'est présentement pas justifiée lorsque L n'est pas majoré par $L_{K_\mathcal{Y}}$.

Conclusion

- L majoré par $L_{\mathcal{K}_\mathcal{Y}} \Rightarrow L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$.
- Dans ce cas, l'approche par régression est justifiée : deux bornes sur le risque sont proposées.
- Le minimiseur de la 1ière borne, SORR, est l'estimateur des moindres carrés (régularisé) étudié par Cortes, Mohri, et Weston (2007).
- Le minimiseur de la 2e borne, SOSC, est nouveau.
- L'approche par régression n'est présentement pas justifiée lorsque L n'est pas majoré par $L_{\mathcal{K}_\mathcal{Y}}$.

Conclusion

- L majoré par $L_{\mathcal{K}_\mathcal{Y}} \Rightarrow L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$.
- Dans ce cas, l'approche par régression est justifiée : deux bornes sur le risque sont proposées.
- Le minimiseur de la 1ière borne, SORR, est l'estimateur des moindres carrés (régularisé) étudié par Cortes, Mohri, et Weston (2007).
- Le minimiseur de la 2e borne, SOSC, est nouveau.
- L'approche par régression n'est présentement pas justifiée lorsque L n'est pas majoré par $L_{K_\mathcal{Y}}$.

Conclusion

- L majoré par $L_{\mathcal{K}_\mathcal{Y}} \Rightarrow L(y_{\mathbf{w}}(x), y) \leq 2\|Y(y) - \mathbf{W}X(x)\|^2$.
- Dans ce cas, l'approche par régression est justifiée : deux bornes sur le risque sont proposées.
- Le minimiseur de la 1ière borne, SORR, est l'estimateur des moindres carrés (régularisé) étudié par Cortes, Mohri, et Weston (2007).
- Le minimiseur de la 2e borne, SOSC, est nouveau.
- L'approche par régression n'est présentement pas justifiée lorsque L n'est pas majoré par $L_{K_\mathcal{Y}}$.

Merci !

Related Work

Canponnetto and De Vito (2007) have established convergence rates of the RLS estimator to $\inf_{f \in \mathcal{H}} R(f)$ that depends on the complexity of the regression function and the effective dimension of $\mathcal{H}$.

In contrast, we provide bounds for any $\mathbf{W}$ in terms of its empirical quadratic risk that does not assume anything about D.

The first PAC-Bayes bound is a uniform risk bound on the *prediction risk* of $\mathbf{W}$ that depends on its empirical quadratic risk.

The PAC-Bayes sample compression bound is a uniform risk bound on the quadratic risk of $\mathbf{W}$ that depends on its empirical quadratic risk and a ℓ_∞ (or ℓ_1) constraint on matrix $\mathbf{A}$.