Comparaison d'algorithmes d'apprentissage et combinaison de modèles.

Alexandre Lacoste

March 22, 2013

Tutoriel sur l'apprentissage bayesien.

Comparaison d'algorithmes et Combinaison de modèles

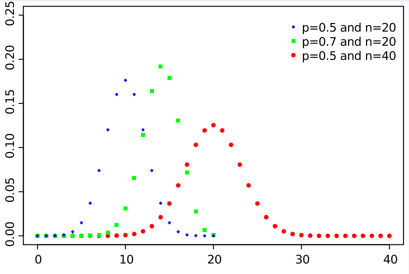
Alexandre Lacoste

March 22, 2013

Distribution Binomiale

Probabilité d'observer k fois un évènement de probabilité q après n essais.

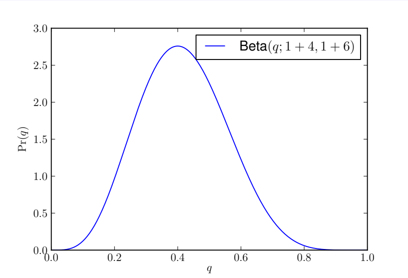
$$\Pr(\#\text{pile} = k | q, n)$$



Distribution Beta

Probabilité que "*la probabilité d'observer pile soit q*" étant donné que nous avons observé k fois "pile" après n tentatives. (inverse de la binomiale).

$$\Pr(q \mid \#\text{pile} = k, n)$$
$$= \text{Beta}(q; k+1, n-k+1)$$



Marginalisation

$$\Pr(A) = \sum_B \Pr(A, B)$$

Factorisation

$$\begin{aligned}\Pr(A, B) &= \Pr(A)\Pr(B|A) \\ &= \Pr(B)\Pr(A|B)\end{aligned}$$

Théorème de Bayes

$$\Pr(B|A) = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}$$

Marginalisation

$$\Pr(A) = \sum_B \Pr(A, B)$$

Factorisation

$$\begin{aligned}\Pr(A, B) &= \Pr(A)\Pr(B|A) \\ &= \Pr(B)\Pr(A|B)\end{aligned}$$

Théorème de Bayes

$$\Pr(B|A) = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}$$

Probabilités 101

Marginalisation

$$\Pr(A) = \sum_B \Pr(A, B)$$

Factorisation

$$\begin{aligned}\Pr(A, B) &= \Pr(A)\Pr(B|A) \\ &= \Pr(B)\Pr(A|B)\end{aligned}$$

Théorème de Bayes

$$\Pr(B|A) = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}$$

Interprétons le résultat

X: Ensemble d'observations

θ: Paramètre

Théorème de Bayes

$$\Pr(\theta|X) = \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X)}$$

$\Pr(\theta)$: Distribution *a priori* sur θ.

$\Pr(\theta|X)$: Distribution *a posteriori* sur θ.

$\Pr(X|\theta)$: Vraisemblance.

$\Pr(X) = \sum_\theta \Pr(X|\theta)\Pr(\theta)$: Vraisemblance marginale.

Théorème de Bayes avec distributions conjuguées

Si la distribution *a priori* est la distribution conjuguée de la vraisemblance, alors la distribution a posteriori sera aussi la distribution conjuguée mais avec des paramètres dépendant des nouvelles observations.

$$g^c(\theta|X) \propto g(X|\theta)\; g^c(\theta|\{\})$$

Distribution conjuguée

Soit $f(a, b)$, une fonction quelconque,
tel que $\forall a$ et $\forall b$, $f(a, b) >= 0$.

Normalisation

$$Z_b = \int f(a, b) da$$

$$Z_a = \int f(a, b) db$$

Ces deux distributions sont le conjuguée l'une de l'autre

$$g(a|b) \stackrel{\text{def}}{=} \frac{1}{Z_b} f(a, b)$$

$$g^c(b|a) \stackrel{\text{def}}{=} \frac{1}{Z_a} f(a, b)$$

Retournons à notre exemple

Pour un n fixe,

$$\Pr(q|k) \propto \text{Binomial}(k|q)\text{Beta}(q|\alpha_0, \beta_0)$$

après calculs et renormalization, nous avons :

$$\Pr(q|k) = \text{Beta}(q|\alpha_0 + k, \beta_0 + n - k)$$

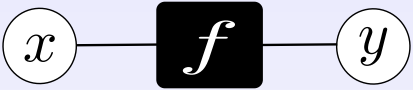
Paramètres de prior

$$\alpha_0 = \beta_0 = 1 \Rightarrow \text{ prior uniforme}$$

Autres distributions

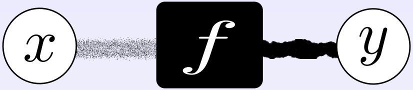
Vraisemblance		**Conjuguée**
Binomial	$\rightarrow$	Beta
Multinomial	$\rightarrow$	Dirichlet
Normal (σ fixe)	$\rightarrow$	Normal
Normal (μ fixe)	$\rightarrow$	Inverse Gamma
Normal	$\rightarrow$	Normal - Inverse Gamma
	$\vdots$	

Apprentissage automatique



- *Deviner* la fonction f à partir d'une collection de paires d'observations $S \stackrel{\text{def}}{=} \{(x_i, y_i)\}_{i=1}^{m}$.
- Par la suite, nous pourrons *Généraliser* sur les x dont nous ignorons la valeur y.

Apprentissage automatique



- *Deviner* la fonction f à partir d'une collection de paires d'observations $S \stackrel{\text{def}}{=} \{(x_i, y_i)\}_{i=1}^{m}$ potentiellement bruités.
- Par la suite, nous pourrons *Généraliser* sur les x dont nous ignorons la valeur y.

Formulation Bayesienne

Cherchons f à partir de S

$$\Pr(f|S) \propto \Pr(S|f)\Pr(f)$$

Modélisons la vraisemblance

$$\begin{aligned}\Pr(S|f) &= \prod_i \Pr(x_i, y_i|f) \\ &= \prod_i \Pr(y_i|x_i, f)\Pr(x_i)\end{aligned}$$

Modèle de bruit

$$\Pr(y_i|x_i, f)$$

Comment faire de nouvelles prédictions

- Notre objectif final est en fait de trouver une réponse y pour un nouveau x.

Maximum a posteriori (MAP) ?

$$f^\star = \operatorname*{argmax}_f \Pr(f|S)$$
$$y = f^\star(x)$$

NON!

- Choisir uniquement le prédicteur le plus probable est équivalent à faire du sur-apprentissage (overfitting).

Comment faire de nouvelles prédictions

- Notre objectif final est en fait de trouver une réponse y pour un nouveau x.

Maximum a posteriori (MAP) ?

$$f^\star = \underset{f}{\operatorname{argmax}} \Pr(f|S)$$

$$y = f^\star(x)$$

NON!

- Choisir uniquement le prédicteur le plus probable est équivalent à faire du sur-apprentissage (overfitting).

Comment faire de nouvelles prédictions (prise 2)

- Comme nous ne connaissons pas explicitement f, il faut marginaliser.

Marginalisons f

$$
\begin{aligned}
\Pr(y|x,S) &= \sum_f \Pr(y,f|x,S) \\
&= \sum_f \Pr(y|x,f)\Pr(f|S)
\end{aligned}
$$

- Nous interrogeons l'opinion de toutes les fonctions pour obtenir un distribution sur les y.
- Pas de sur-apprentissage ☺

Théorie de la Décision Bayesienne

- Pour pouvoir se prononcer sur une valeur de y, il faut connaître le coût associé à nos actions

Fonction de perte $\mathcal{L}$

$$l_i = \mathcal{L}(y_i, f(x_i))$$

Le coût associé à répondre $f(x_i)$ lorsque la vraie réponse était y_i

Décision optimale

$$\begin{aligned} y^\star &= \operatorname*{argmin}_y \sum_{\widehat{y}} \Pr(\widehat{y}|S,x)\mathcal{L}(y,\widehat{y}) \\ &= \operatorname*{argmin}_y \mathrm{Risk}(y|S,x) \end{aligned}$$

Résumons

Définir un modèle de bruit et un prior

$$\Pr(y_i|x_i, f), \quad \Pr(f)$$

Cherchons f à partir de S

$$\Pr(f|S) \propto \Pr(f) \prod_i \Pr(y_i|x_i, f)$$

Marginalisons f

$$\Pr(y|x, S) = \sum_f \Pr(y|x, f) \Pr(f|S)$$

Décision optimale

$$y^\star = \operatorname*{argmin}_y \sum_{\widehat{y}} \Pr(\widehat{y}|S, x) \mathcal{L}(y, \widehat{y})$$

- Apprentissage automatique résolu ?

Pas tout a fait ...

- Pour chaque nouvelle tâche, il faut modéliser le bruit et le prior.
- De manière générale, la marginalisation et normalisation est computationnellement très couteux.

- Apprentissage automatique résolu ?

Pas tout a fait ...

- Pour chaque nouvelle tâche, il faut modéliser le bruit et le prior.
- De manière générale, la marginalisation et normalisation est computationnellement très couteux.

Recommencons les 2 premières étapes

Définir un modèle de bruit et un prior

$$\Pr(y_i|x_i, f), \quad \Pr(f)$$

Cherchons f à partir de S

$$\Pr(f|S) \propto \Pr(f) \prod_i \Pr(y_i|x_i, f)$$

Marginalisons f

$$\Pr(y|x, S) = \sum_f \Pr(y|x, f) \Pr(f|S)$$

Décision optimale

$$y^\star = \underset{y}{\operatorname{argmin}} \sum_{\widehat{y}} \Pr(\widehat{y}|S, x) \mathcal{L}(y, \widehat{y})$$

Recommencons les 2 premières étapes

Apprentissage Agnostic Bayes !

Utilisons $\mathcal{L}$ pour éviter d'avoir a spécifier un modèle de bruit

Marginalisons f

$$\Pr(y|x,S) = \sum_f \Pr(y|x,f)\Pr(f|S)$$

Décision optimale

$$y^\star = \operatorname*{argmin}_y \sum_{\widehat{y}} \Pr(\widehat{y}|S,x)\mathcal{L}(y,\widehat{y})$$

Apprentissage agnostique

$\mathcal{F}$: L'ensemble de tous les prédicteurs f potentiels.

Tâche D : La *vraie* probabilité d'observer (x, y) (c.à.d.: $\Pr(x, y)$).

$R_D(f)$: $\mathop{\mathbf{E}}_{x,y\sim D} \mathcal{L}\left(y, f(x)\right)$ (risque).

$\mathcal{L}$: Fonction de perte.

Objectif idéal si on connaissais D

$$f^\star = \operatorname*{argmin}_{f\in\mathcal{F}} R_D(f)$$

- Mais, nous ne connais pas explicitement D, il faudra travailler avec S.

Exemple minimal

Suppositions

- perte zero-un : $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- $|\mathcal{F}| = 2$
- $|S| = 6$

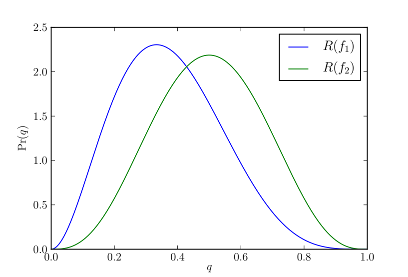
	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1
$R_S(f)$	2/6	3/6

Exemple minimal

Suppositions

- perte zero-un : $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- $|\mathcal{F}| = 2$
- $|S| = 6$

	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1
$R_S(f)$	2/6	3/6



Il faut tenir compte des correlations !!

f_1 et f_2 sont tous les deux testés sur les mêmes données donc les séquences de loss sont corrélés entre elles.

	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1
$R_S(f)$	2/6	3/6

4 évènements différents

- 00
- 01
- 10
- 11

Seulement 2 évènements importants

- 01
- 10

Theorem 4.1. *Let* $\alpha_h \stackrel{\text{def}}{=} \alpha'_h + k_h$, $\alpha_g \stackrel{\text{def}}{=} \alpha'_g + k_g$ *and* $\overline{\alpha} \stackrel{\text{def}}{=} \overline{\alpha}' + \overline{k}$, *where* $\alpha'_g > 0$, $\alpha'_h > 0$, $\overline{\alpha}' > 0$, *then*

$$\begin{aligned}
&\Pr\left(h \stackrel{\mathcal{D}}{\succ} g\right) \\
&= \int_0^1 \int_0^{\frac{1-\overline{p}}{2}} D\left(p_g, p_h, \overline{p}\ ;\ \alpha_g, \alpha_h, \overline{\alpha}\right)\ dp_h\ d\overline{p} \\
&= B_c\left(\tfrac{1}{2}; \alpha'_h + k_h, \alpha'_g + k_g\right)
\end{aligned}$$

Proof. The first equality follows from the explanations above. Now, using $C \stackrel{\text{def}}{=} \frac{\Gamma(\alpha_h+\alpha_g+\overline{\alpha})}{\Gamma(\alpha_h)\Gamma(\alpha_g)\Gamma(\overline{\alpha})}$, $\gamma \stackrel{\text{def}}{=} 1-\overline{p}$ and $z \stackrel{\text{def}}{=} \frac{p_h}{\gamma}$, we have :

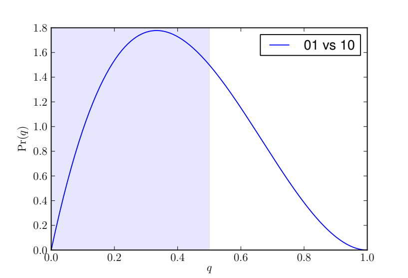
$$\begin{aligned}
&\int_0^1 \int_0^{\frac{1-\overline{p}}{2}} D\left(p_g, p_h, \overline{p}\ ;\ \alpha_g, \alpha_h, \overline{\alpha}\right)\ dp_h\ d\overline{p} \\
&= C \int_0^1 \overline{p}^{\overline{\alpha}-1} \int_0^{\frac{1-\overline{p}}{2}} p_h^{\alpha_h-1} \left(1-\overline{p}-p_h\right)^{\alpha_g-1}\ dp_h\ d\overline{p} \\
&= C \int_0^1 \overline{p}^{\overline{\alpha}-1} \int_0^{\frac{1}{2}} (\gamma z)^{\alpha_h-1} (\gamma-\gamma z)^{\alpha_g-1}\ \gamma\ dz\ d\overline{p} \\
&= C \int_0^1 \overline{p}^{\overline{\alpha}-1} \gamma^{\alpha_h+\alpha_g-1}\ d\overline{p} \int_0^{\frac{1}{2}} z^{\alpha_h-1} (1-z)^{\alpha_g-1}\ dz \\
&= \frac{\Gamma(\alpha_h+\alpha_g)}{\Gamma(\alpha_h)\Gamma(\alpha_g)} \int_0^{\frac{1}{2}} z^{\alpha_h-1} (1-z)^{\alpha_g-1}\ dz \\
&\stackrel{\text{def}}{=} B_c\left(\tfrac{1}{2}; \alpha'_h + k_h, \alpha'_g + k_g\right)
\end{aligned}$$

$\square$

Probabilité que f_1 soit meilleur que f_2

	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1

$\mathrm{Beta}(q; 1+2, 1+1)$



Cumulative de la Beta

$$\Pr\left(R_D\left(f_1\right) < R_D\left(f_2\right)\right)$$

$$= \int_{q=0}^{\frac{1}{2}} \mathrm{Beta}(q; 1+k_{01}, 1+k_{10})dq$$

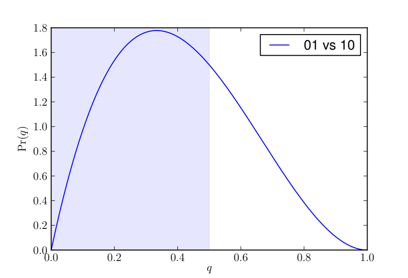
$\Pr(f = f^\star | S)$

- $\Pr(f_1 = f^\star | S) = 0.69$
- $\Pr(f_2 = f^\star | S) = 0.31$

Probabilité que f_1 soit meilleur que f_2

	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1

$\mathrm{Beta}(q; 1+2, 1+1)$



Cumulative de la Beta

$$\Pr\left(R_D\left(f_1\right) < R_D\left(f_2\right)\right)$$

$$= \int_{q=0}^{\frac{1}{2}} \mathrm{Beta}(q; 1+k_{01}, 1+k_{10}) dq$$

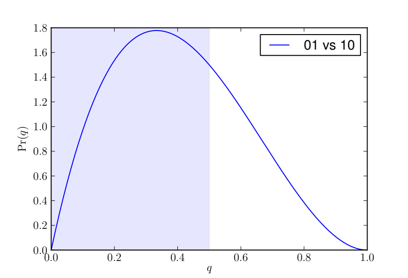
$\Pr(f = f^{\star}|S)$

- $\Pr(f_1 = f^{\star}|S) = 0.69$
- $\Pr(f_2 = f^{\star}|S) = 0.31$

Probabilité que f_1 soit meilleur que f_2

	f_1	f_2
x_1, y_1	0	0
x_2, y_2	0	1
x_3, y_3	1	0
x_4, y_4	0	0
x_5, y_5	1	1
x_6, y_6	0	1

$\mathrm{Beta}(q; 1+2, 1+1)$



Cumulative de la Beta

$$\Pr\left(R_D\left(f_1\right) < R_D\left(f_2\right)\right)$$

$$= \int_{q=0}^{\frac{1}{2}} \mathrm{Beta}(q; 1+k_{01}, 1+k_{10}) dq$$

$\Pr(f = f^\star | S)$

- $\Pr(f_1 = f^\star | S) = 0.69$
- $\Pr(f_2 = f^\star | S) = 0.31$

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

- Nous pouvons obtenir un posterieur *Agnostic Bayes* pour $|\mathcal{F}| = 2$ et $\mathcal{L}(y, y') = \mathbf{1}_{y \neq y'}$
- Pouvons nous obtenir le même résultat pour $|\mathcal{F}| > 2$?
- Et pour n'importe quel type de fonction $\mathcal{L}$?

Oui ☺

- Mais je ne vais pas vous dire comment :p
- L'algorithme a une complexité algorithmique de $O(|S||\mathcal{F}|)$.

Récapitulons ...

Objectif classique

$$\Pr(f = f^{\mapsto} | S) \propto \Pr(f) \prod_i \Pr(y_i | x_i, f)$$

Objectif agnostique

$$f^\star = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R(f)$$

$$\Pr(f = f^\star) = \Pr\left(R(f) < R(f')\ , \forall f' \neq f\right)$$

Marginalisons f

$$\Pr(y|x, S) = \sum_f \Pr(y|x, f)\Pr(f|S)$$

Décision optimale

$$y^\star = \underset{y}{\operatorname{argmin}} \sum_{\widehat{y}} \Pr(\widehat{y}|S, x)\mathcal{L}(y, \widehat{y})$$

Récapitulons ...

Objectif classique

$$\Pr(f = f^{\mapsto} | S) \propto \Pr(f) \prod_i \Pr(y_i | x_i, f)$$

Objectif agnostique

$$f^\star = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R(f)$$

$$\Pr(f = f^\star) = \Pr\Big(R(f) < R(f')\ , \forall f' \neq f\Big)$$

Marginalisons f

$$\Pr(y|x, S) = \sum_f \Pr(y|x, f) \Pr(f|S)$$

Décision optimale

$$y^\star = \underset{y}{\operatorname{argmin}} \sum_{\widehat{y}} \Pr(\widehat{y}|S, x) \mathcal{L}(y, \widehat{y})$$

Récapitulons ...

Objectif classique

$$\Pr(f = f^{\mapsto} | S)$$
$$\propto \Pr(f) \prod_i \Pr(y_i | x_i, f)$$

Objectif agnostique

$$f^{\star} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R(f)$$

$$\Pr(f = f^{\star})$$
$$= \Pr\Big(R(f) < R(f') \;, \forall f' \neq f \Big)$$

Marginalisons f

$$\Pr(y | x, S) = \sum_f \Pr(y | x, f) \Pr(f | S)$$

Décision optimale

$$y^{\star} = \underset{y}{\operatorname{argmin}} \sum_{\widehat{y}} \Pr(\widehat{y} | S, x) \mathcal{L}(y, \widehat{y})$$

Récapitulons ...

Objectif classique

$$\Pr(f = f^{\mapsto} | S)$$
$$\propto \Pr(f) \prod_i \Pr(y_i | x_i, f)$$

Objectif agnostique

$$f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$$

$$\Pr(f = f^\star)$$
$$= \Pr\Big(R(f) < R(f')\ , \forall f' \neq f\Big)$$

Marginalisons f

$$\Pr(y|x, S) = \sum_f \Pr(y|x, f) \Pr(f|S)$$

Décision optimale

$$y^\star = \operatorname*{argmin}_y \sum_{\widehat{y}} \Pr(\widehat{y}|S, x) \mathcal{L}(y, \widehat{y})$$

Limitations de l'apprentissage agnostique Bayes

- Pour le moment, nous sommes limités à $|\mathcal{F}| < \infty$

☹

- Je crois qu'il est possible de généraliser à des classes infini non dénombrables.
- Mais, même avec une classe de taille fini, nous pouvons avoir des applications utiles!

Limitations de l'apprentissage agnostique Bayes

- Pour le moment, nous sommes limités à $|\mathcal{F}| < \infty$

$$\frown$$

- Je crois qu'il est possible de généraliser à des classes infini non dénombrables.
- Mais, même avec une classe de taille fini, nous pouvons avoir des applications utiles!

Limitations de l'apprentissage agnostique Bayes

- Pour le moment, nous sommes limités à $|\mathcal{F}| < \infty$

$\frownie$

- Je crois qu'il est possible de généraliser à des classes infini non dénombrables.
- Mais, même avec une classe de taille fini, nous pouvons avoir des applications utiles!

Validation croisée agnostique Bayes

- En validation croisée, nous entraînons et évaluons un nombre fini de modèles

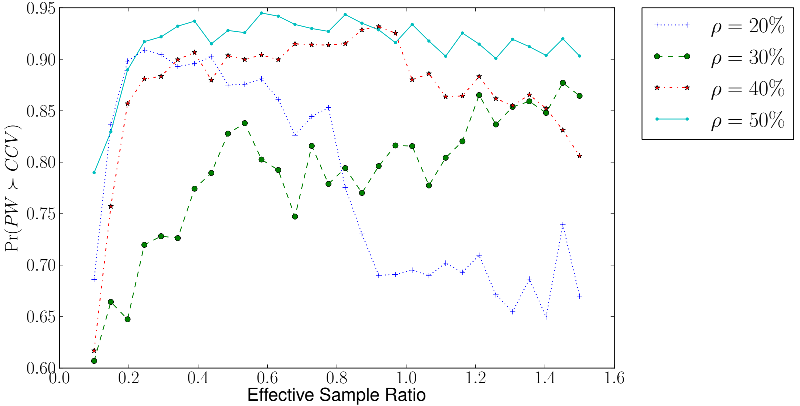
sur-apprentissage

Choisir le *meilleur* modèle en validation croisée
$\Rightarrow$ sur-apprentissage

Combinaison de modèles

En utilisant la marginalisation du posterieur agnostique, nous pouvons combiner différent algorithmes d'apprentissgae avec différents hyperparmètres.

Réultats sur la validation croisée



Des questions ?