Modèles génératifs et discriminatifs

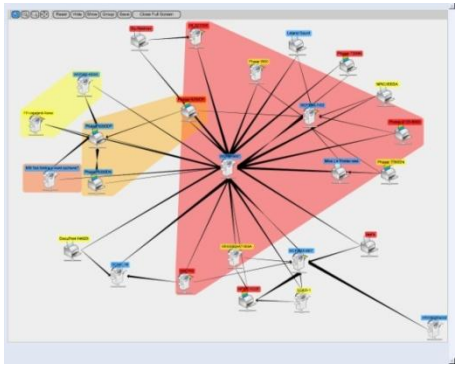
Guillaume Bouchard
Xerox Research Center Europe, Grenoble

Journées MAS 2010
31 août - 3 septembre 2010

Motivation

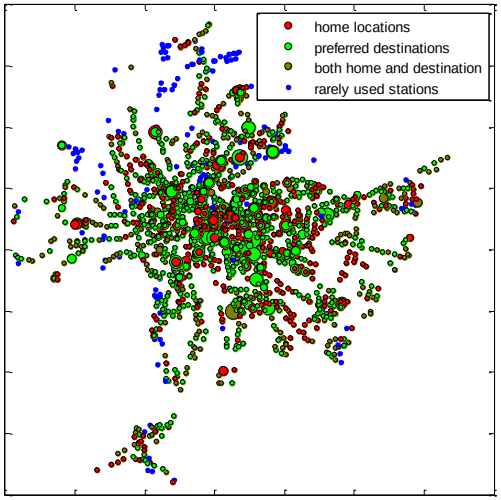
- Nombreux problemes de classification peu de données labelisées

Détection et localisation de pannes d'imprimante



Détection automatique de paraphrases dans des données textuelles

Profilage d'utilisateurs de transport en commun



Classification supervisée

- Entrées - Sorties
  - $\mathrm{x} \in \mathcal{X}$
  - $\mathrm{y} \in \{1, \cdots, K\}$
- Données d'apprentissage: (x_i, y_i), i=1,…,n i.i..d de poi p*

Approche discriminative	Approche generative
Famille de probabilités conditionnelles $\{p_\theta(y \mid x), \theta \in \Theta_D\}$	Famille de probabilités jointes $\{p_\theta(y,x), \theta \in \Theta_G\}$

- Objectif
  - estimer θ de manière à minimiser le cout de la décision
  - On utilise le cout $\mathrm{E}^*[\text{-}\log(p_\theta(y \mid x)]$

Cadre théorique

- Famille exponentielle

$$p_\theta(x,y) = \exp\{\phi(x,y)^\top\theta - A(\theta)\}, \quad A(\theta) = \log\int_{\mathcal{X}}\int_{\mathcal{Y}}\exp\{\phi(x,y)^\top\theta\}dydx,$$

$$p_\theta(y \mid x) = \exp\{\phi(x,y)^\top\theta - A(\theta;x)\}, \quad A(\theta;x) = \log\int_{\mathcal{Y}}\exp\{\phi(x,y)^\top\theta\}dy.$$

- $\phi(x,y)\in R^{\mathrm{d}}$ est le vecteur de statistiques suffisantes
- Estimateurs génératifs et discriminatifs

$$\hat{\theta}_n^{\mathrm{gen}} \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_\theta G_n(\theta), \quad G_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(x,y),$$

$$\hat{\theta}_n^{\mathrm{dis}} \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_\theta D_n(\theta), \quad D_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(y \mid x).$$

Modele bien spécifié

- Si la famille parametrique contient la densité des données, i.e.

$$p^*(x,y) = p_{\theta_\infty}(x,y)$$

$\Rightarrow$ Alors, le modele génératif est asymptotiquement optimal

- Il est meme strictement meilleur [Liang&Jordan,2008]

$$\mathcal{L}(\hat{\theta}_n^{\text{gen}}) \leq \mathcal{L}(\hat{\theta}_n^{\text{dis}}) - \frac{c}{n} + O_p(n^{-\frac{3}{2}})$$

(Meilleure information de Fisher en modélisant $p(x)$)

Modele mal spécifié

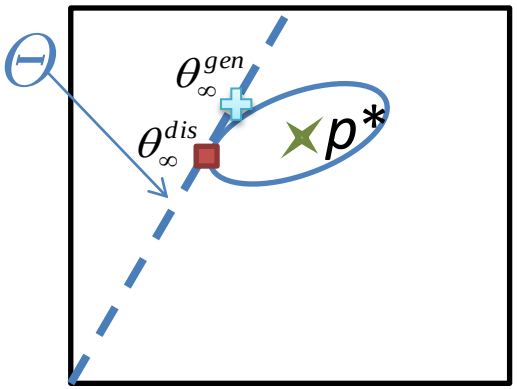
- En général, p^* n'appartient pas a une famille parametrique finie

➔l'estimateur discriminatif meilleur asymptotiquement

Exemple (1)

Distribution p^* des données (2 paramètres)

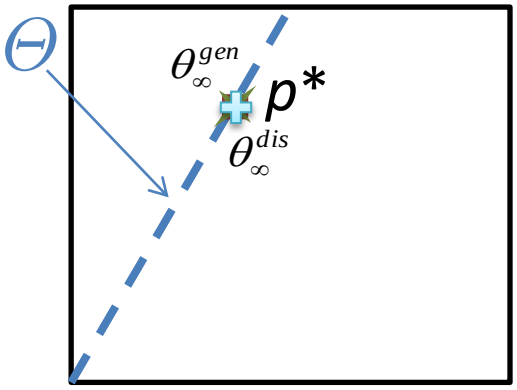
- $Y \sim \mathcal{B}(0.5)$
- $X|Y=1 \sim \mathcal{B}(p_1^*)$
- $X|Y=0 \sim \mathcal{B}(p_0^*)$



$p_0^* \neq 2p_1^*$: modèle mal spécifié

Famille paramétrique approchée (1 paramètre)

- $Y \sim B(0.5)$
- $X|Y=1 \sim B(\theta)$
- $X|Y=0 \sim B(2\theta)$
- $\theta \in \Theta = [0.5]$

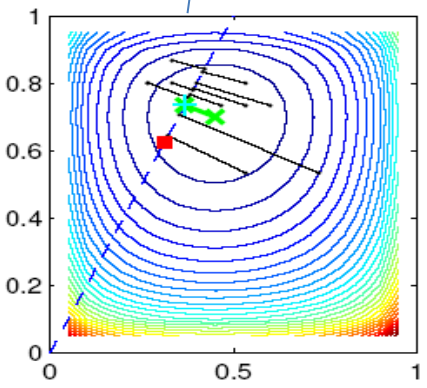


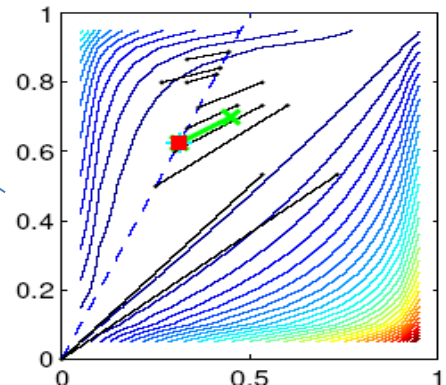
$p_2^* = 2p_1^*$: modèle mal spécifié

Exemple (2)

- $p_1^* = 0.45, p_0^* = 0.70$
- $Biais^2 = E^*\left[-\log(p_{\theta_\infty}(y \mid x))\right] - E^*\left[-\log(p^*(y \mid x))\right]$
- $Variance = E^*\left[-\log(p_{\hat{\theta}_n}(y \mid x))\right] - E^*\left[-\log(p_{\theta_\infty}(y \mid x))\right]$
- n=30 samples

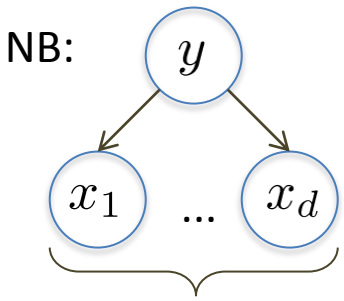
	$Biais^2$	$Variance$	Total
Modèle saturé	0	0.2382	0.2382
Estimateur discriminatif	0.1271	0.1354	0.2625
Estimateur génératif	0.2249	0.0184	0.2433



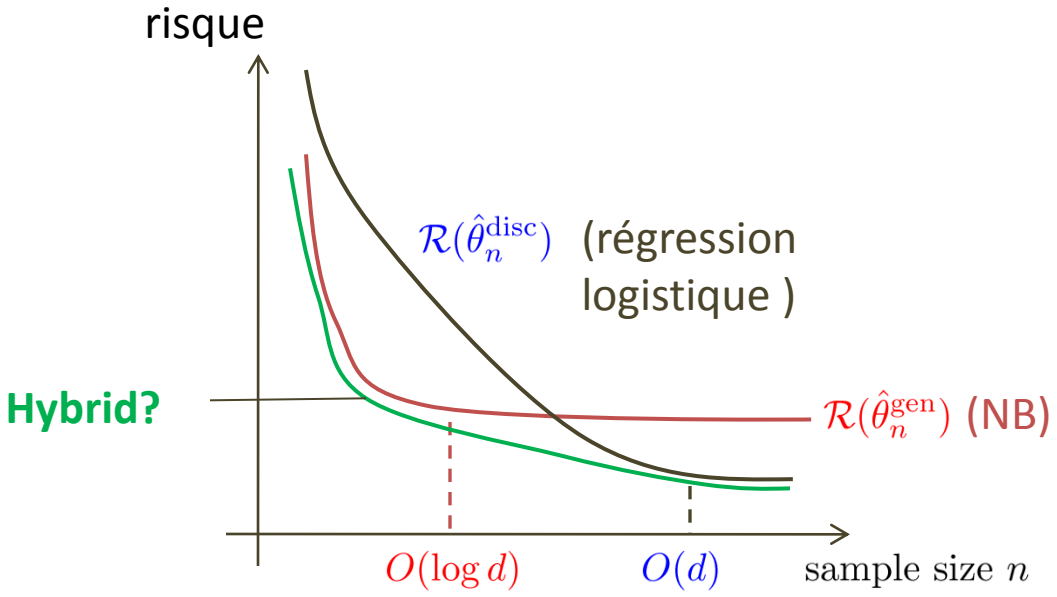


Motivation pour des modèles hybrides

- [Ng & Jordan NIPS 02]: Naive Bayes vs. régression logistique pour la classification binaire linéaire



d covariables indépendantes dans chaque classe



Estimation hybride générative-discriminative

- Interpolation entre les estimateurs génératifs et discriminatif:

$$\hat{\theta}_n^{\lambda} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, D_n(\theta) + \frac{\lambda}{n} G_n(\theta)$$

- Peut être interprété comme un « lissage » aléatoire
- Intuition pour une famille « presque correcte »:

  réduction de variance > accroissement du biais

  ➔gain de performance prédictive

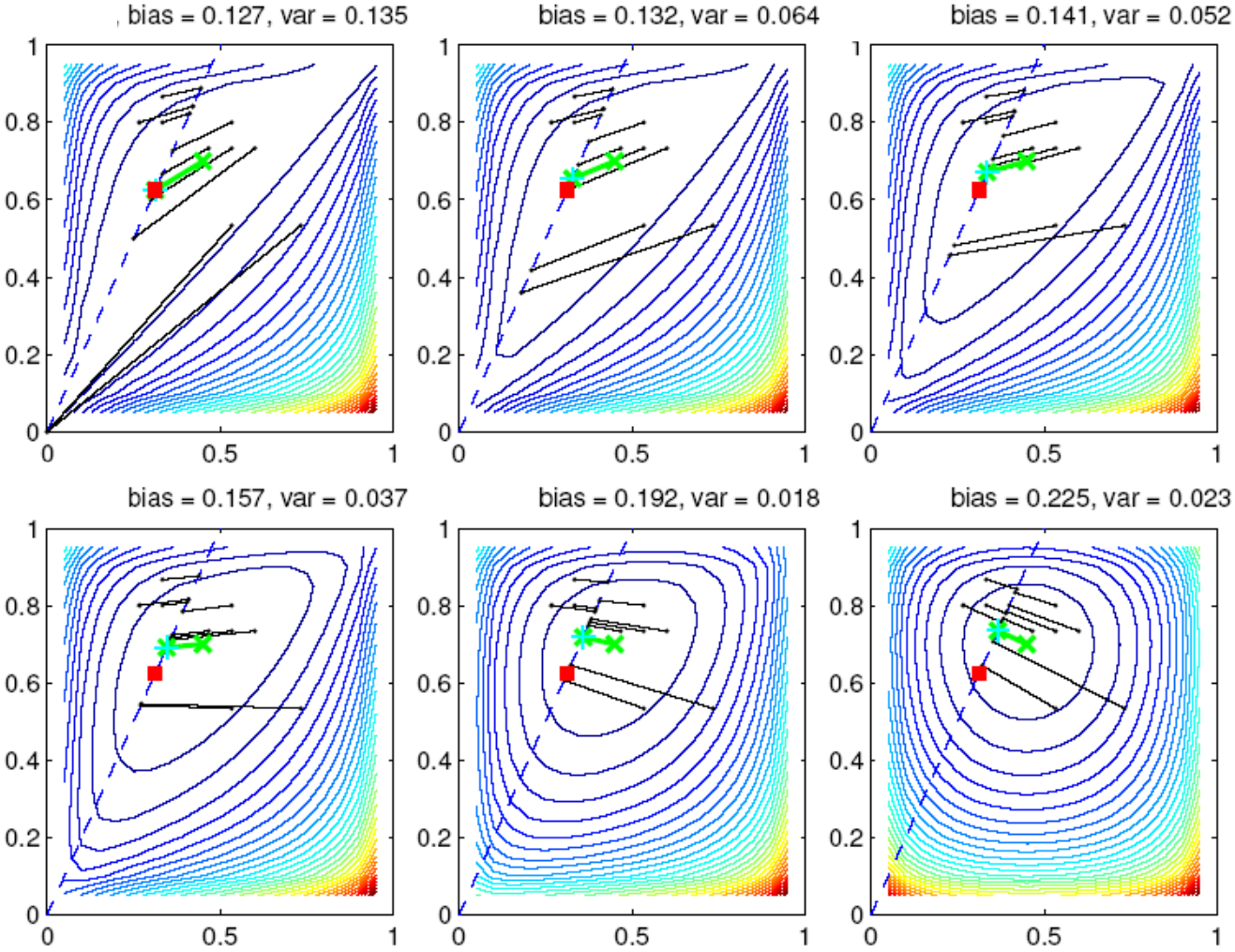
Exemple (4)

- Equilibre le biais et la variance d'estimation

λ	bias^2 $L(\theta^*_\lambda) - L(\theta^*_C)$	variance $E\left[L(\hat{\theta}_\lambda) - L(\theta^*_\lambda)\right]$	bias^2+variance $E\left[L(\hat{\theta}_\lambda) - L(\theta^*_C)\right]$
Saturated model	0	0.2382	0.2382
0 (Discriminative)	0.1271	0.1354	0.2625
0.05	0.1324	0.0640	0.1964
0.10	0.1409	0.0519	0.1928
0.25	0.1572	0.0369	0.1942
0.50	0.1923	0.0182	0.2105
∞ (Generative)	0.2249	0.0184	0.2433

Exemple(5)

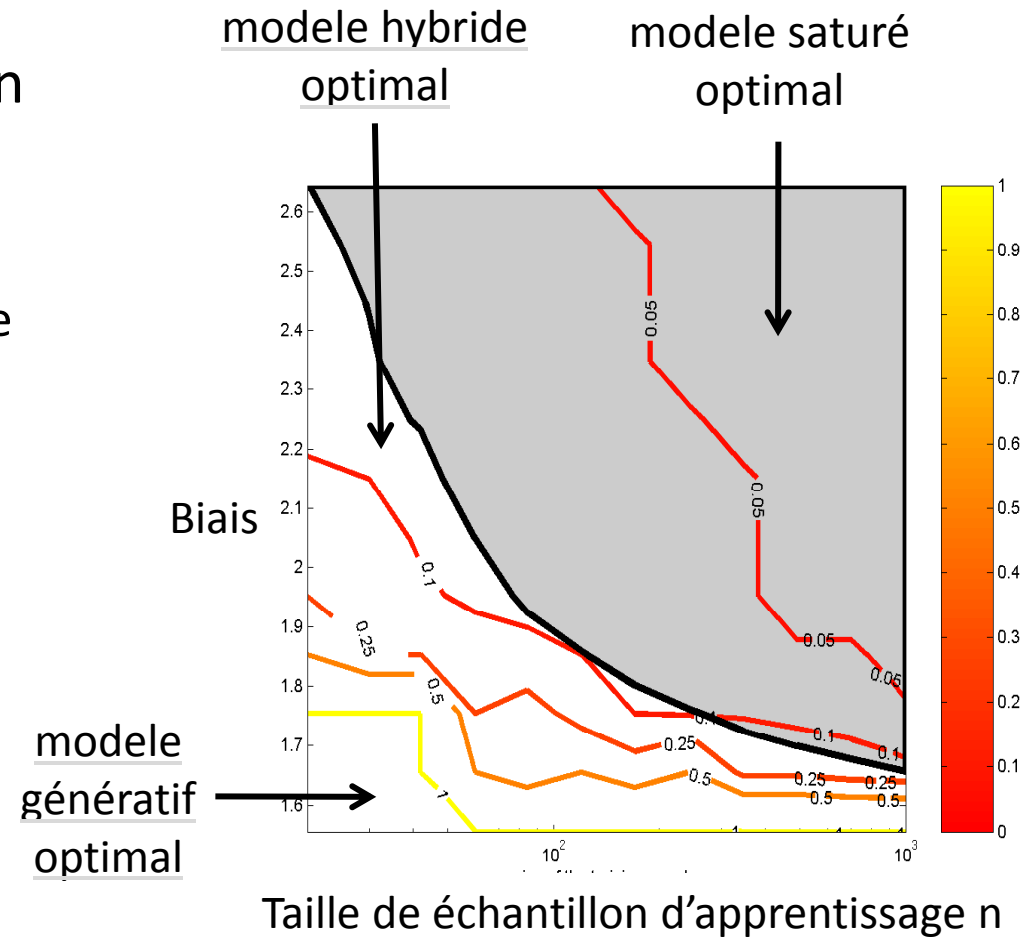
$\lambda = 0$ (discriminatif) $\lambda = 0.05$ $\lambda = 0.1$



$\lambda = 0.25$ $\lambda = 0.5$ $\lambda = \infty$ (génératif)

Choix du paramètre de régularisation

- Quelle est la régularisation optimale?
  - Validation croisée ?
  - Critère de sélection de type AIC/BIC?



Régularisation optimale

- Cas général [Liang et al., 2010]

$$\lambda^* = \frac{\mathrm{tr}\{\mathcal{I}_{\ell\ell}\mathcal{V}_{\mathrm{x}}^{-1}\mathcal{V}\mathcal{V}_{\mathrm{x}}^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{\mathrm{xy}}) - \mathrm{tr}\{\mathcal{I}_{\ell r}\mathcal{V}_{\mathrm{x}}^{-1}\}}{\mathrm{tr}\{(\mu - \mu_{\mathrm{xy}})^{\otimes}\mathcal{V}_{\mathrm{x}}^{-1}\}}$$

$$\begin{aligned}
\mu_{\mathrm{xy}} &\stackrel{\mathrm{def}}{=} \mathbb{E}_{p^*(X,Y)}[\phi(X,Y)], \\
\mu_{\mathrm{x}} &\stackrel{\mathrm{def}}{=} \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X,Y)], \\
\mu &\stackrel{\mathrm{def}}{=} \mathbb{E}_{p_{\theta_\infty}(X,Y)}[\phi(X,Y)], \\
\mathcal{V}_{\mathrm{x}} &\stackrel{\mathrm{def}}{=} \mathbb{E}_{p^*(X)}[\mathbb{V}_{p_{\theta_\infty}(Y|X)}[\phi(X,Y)]], \\
\mathcal{V} &\stackrel{\mathrm{def}}{=} \mathbb{V}_{p_{\theta_\infty}(X,Y)}[\phi(X,Y)]].
\end{aligned}$$

$$\begin{aligned}
\mathcal{I}_{\ell\ell} &\stackrel{\mathrm{def}}{=} \mathbb{E}[\dot{\ell}(Z;\theta_\infty)^{\otimes}] \\
\mathcal{I}_{\ell r}(\lambda) &\stackrel{\mathrm{def}}{=} \lim_{n\to\infty} n^2 \cdot \mathbb{E}[\dot{L}_n(\theta_\infty)\dot{R}_n(\lambda,\theta_\infty)^\top]
\end{aligned}$$

- Cas ou le modèle conditionnel est correct $p^*(y \mid x) = p_{\theta_\infty}(y \mid x)$

$$\lambda^* = \frac{\mathrm{tr}\{(\mathcal{V} - \mathcal{V}_{\mathrm{x}})\mathcal{V}_{\mathrm{x}}^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{\mathrm{xy}})}{\mathrm{tr}\{(\mu - \mu_{\mathrm{xy}})^{\otimes}\mathcal{V}_{\mathrm{x}}^{-1}\}}$$

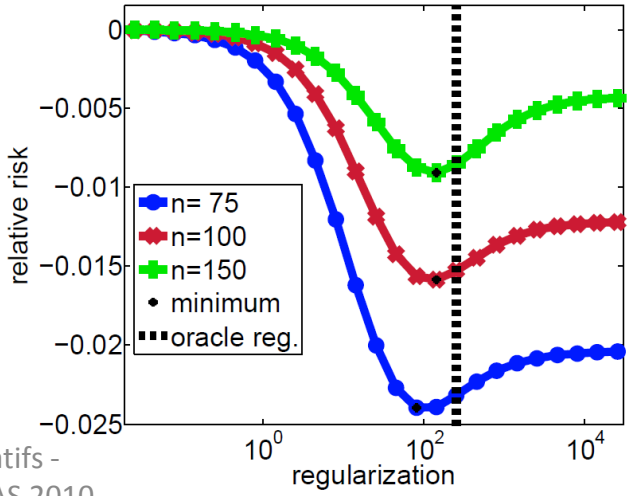
Simulation

$y \in \{0,1\}$, $x \in \{0,1\}^{10}$ $\quad \phi(x,y) = (\mathbb{I}[y=0]x^\top, \mathbb{I}[y=1]x^\top)^\top$

$$\theta_\infty = (\tfrac{1}{10}, \cdots, \tfrac{1}{10}, \tfrac{3}{10}, \cdots, \tfrac{3}{10})^\top$$

$$p^*(x,y) = (1-\varepsilon)p_{\theta_\infty}(x,y) + \varepsilon p_{\theta_\infty}(y)p_{\theta_\infty}(x_1 \mid y)\mathbb{I}[x_1 = \cdots = x_k]$$

Misspecification	$\mathrm{tr}\{\mathcal{I}_{\ell\ell}\mathcal{V}_{\mathrm{x}}^{-1}\mathcal{V}\mathcal{V}_{\mathrm{x}}^{-1}\}$	$2\mathcal{B}^\top(\mu-\mu_{\mathrm{xy}})$	$\mathrm{tr}\{(\mu-\mu_{\mathrm{xy}})^{\otimes}\mathcal{V}_{\mathrm{x}}^{-1}\}$	λ^*	$\mathbb{L}(\lambda^*)$
0%	5	0	0	∞	-0.65
5%	5.38	-0.073	0.00098	310	-48
50%	13.8	-1.0	0.034	230	-808



Expériences numériques

- modèle génératif
  - Indépendance des covariables dans chaque groupe (Modèle Bayésien Naïf)
  - Variables continues Gaussiennes
  - Variables discrètes binarisées
- L'estimation discriminative est équivalente a la régression logistique
- Apprentissage par descente de gradient
- Données de benchmark issues de la base UCI*
- Paramètre de régularisation choisi par validation croisée d'ordre 10
- 50 simulations des jeux de test et d'apprentissage

*http://archive.ics.uci.edu/ml/

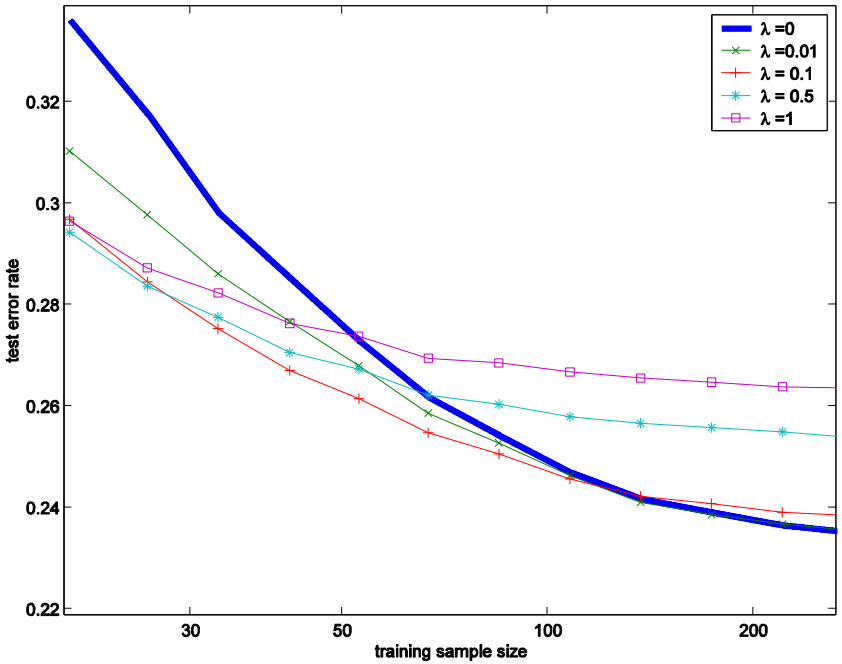
Résultats

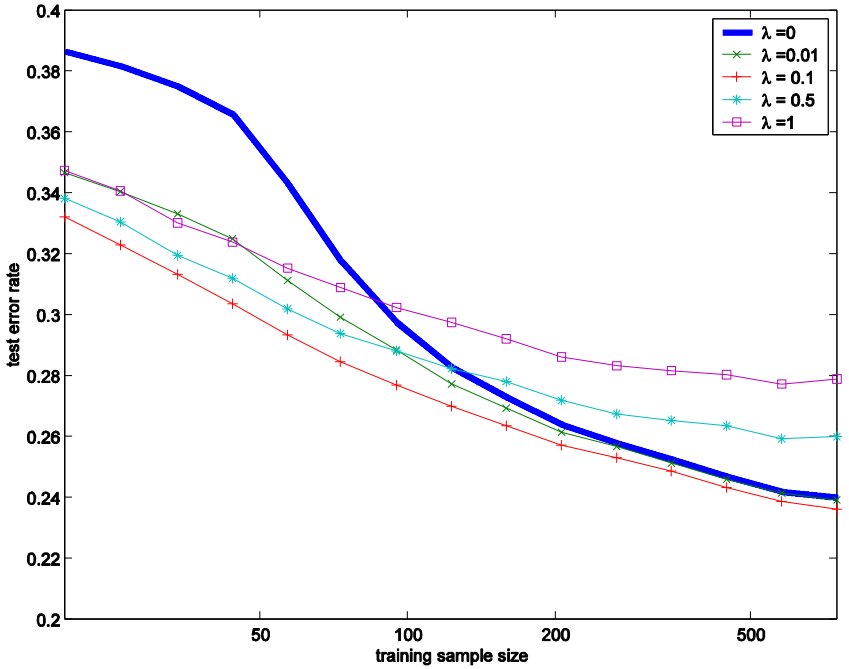
dataset	d	n	err_{DISC}	$\text{err}_{\hat{\lambda}}$	err_{GEN}	$\hat{\lambda}$
abalone	8	99	22.9 ± 1.5	22.1 ± 1.1	22.7 ± 0.34	0.012
contraceptive method choice	9	475	37.3 ± 1.5	37.3 ± 1.5	40.3 ± 2.2	0.01
ionosphere	34	97	16.8 ± 3.2	13.2 ± 1.5	20.1 ± 5.6	0.103
optdigits	64	48	1.65 ± 1.2	1.52 ± 1.1	1.73 ± 1.2	0.174
pageblocks	10	60	4.21 ± 2.1	3.11 ± 0.73	3.69 ± 1.3	0.215
spam	57	420	11.8 ± 1.3	10 ± 0.96	10.8 ± 0.78	0.262
australian	14	142	15.3 ± 1.6	13.6 ± 0.97	13.5 ± 0.9	0.54
diabetes	8	45	27.7 ± 3.1	27 ± 2.4	27.6 ± 2.4	0.588
dna	180	514	5.82 ± 0.83	3.59 ± 0.44	5.64 ± 0.52	0.0302
german	20	73	31.4 ± 2.7	30 ± 2.7	31.4 ± 2.5	0.5
heart	13	179	17.1 ± 3.3	16.7 ± 2.7	16.8 ± 2.9	0.426
letter	16	102	3.12 ± 1.6	2.63 ± 0.84	7.12 ± 1.5	0.00541
shuttle	9	31	9.76 ± 3.9	7.23 ± 2.9	9.79 ± 3.2	0.143
vehicle	18	288	4.37 ± 1.5	3.9 ± 1.5	31.5 ± 4.1	0.001

➔Amélioration significative des taux d'erreur par rapport a la classification générative ***et*** discriminative

Effet de la taille de l'échantillon

- Abalone – 8 dimensions
- German – 20 dimensions





Conclusion

- Améliorations des taux de classification
  - Limitée en théorie (en $1/n^2$) et en pratique (0-5%)
  - Améliorations plus importante attendue pour des problèmes plus structurés (mais estimation difficile)
- Les modèles hybrides permettent essentiellement
  - de garder l'interprétation des paramètres
  - d'utiliser les données non-annotées
- Perspectives
  - Etudes non-asymptotique (borne sur l'erreur empirique)
  - Analyse non-paramétrique (le biais tends vers 0)
- Problèmes ouverts
  - Algorithmes efficaces pour l'estimation discriminative des modèles génératifs
  - Critères de sélection du paramètre de régularisation

Références

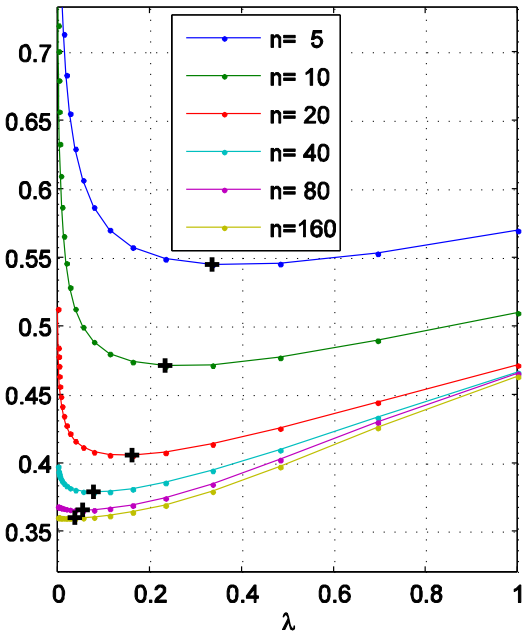
- P. Liang, F. Bach, G. Bouchard, M. I. Jordan. Asymptotically optimal regularization in smooth parametric models. *Advances in Neural Information Processing Systems (NIPS)*, 2009
- G. Bouchard and B. Triggs, The tradeoff between generative and discriminative classifiers. In J. Antoch, editor, Proc. of COMPSTAT'04, 16th Symposium of IASC, volume 16. Physica-Verlag, 2004.
- G. Bouchard, Bias-variance tradeoff in hybrid generative-discriminative models. In proc. of the Sixth International conference on Machine Learning and Applications (ICMLA 07), Cincinnati, Ohio, USA, 13-15 December 2007.
- B. Efron, The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. Journal of the American Statistical Association, 70(352), 892—898, 1975.
- P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In International Conference on Machine Learning (ICML), 2008.
- J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In Computer Vision and Pattern Recognition (CVPR), pages 87-94, 2006.
- [Ng 02] A. Y. Ng and M. I. Jordan, On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. In Advances in Neural Information Processing Systems 14, 2002.

Fin

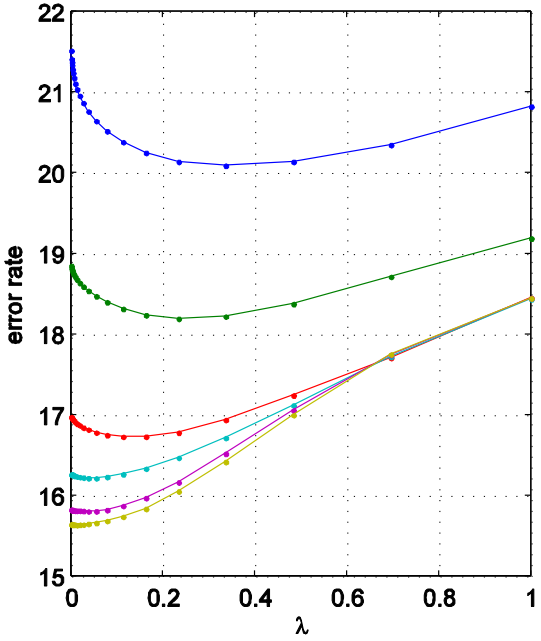
Analyse asymptotique

- Le paramètre de régularisation est inversement proportionnel à la taille de l'échantillon

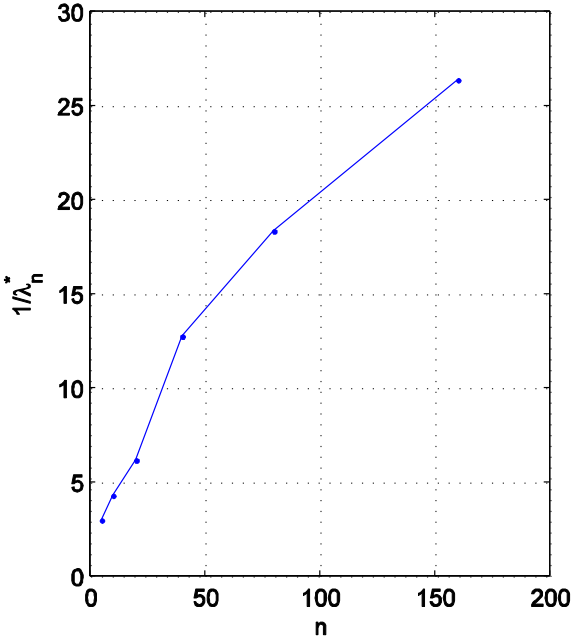
log p(y|x)



error rate



Training sample size
Vs. 1/λ



Full formula

$$\lambda^* = \frac{\mathrm{tr}\{\mathcal{I}_{\ell\ell}\mathcal{V}_{\mathrm{x}}^{-1}\mathcal{V}\mathcal{V}_{\mathrm{x}}^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{\mathrm{xy}}) - \mathrm{tr}\{\mathcal{I}_{\ell r}\mathcal{V}_{\mathrm{x}}^{-1}\}}{\mathrm{tr}\{(\mu - \mu_{\mathrm{xy}})\otimes\mathcal{V}_{\mathrm{x}}^{-1}\}}$$