Classification, codage et entropie

Olivier Catoni

CNRS
DMA

Jeudi 2 octobre 2008

Classification supervisée

On part d'un ensemble de données étiquetées indépendantes

$$(X_i, Y_i)_{i=1}^N \sim \bigotimes_{i=1}^N P_i \in \mathcal{M}_+^1\left[(\mathcal{X} \times \mathcal{Y})^N\right].$$

L'indépendance est imposée par la façon dont les données ont été collectées. L'hypothèse $P_i \equiv P$ peut aussi être garantie de façon réaliste, mais par contre les lois P_i sont en général compliquées, inconnues et portent sur un espace d'état $\mathcal{X}$ de très grande dimension. On peut aussi tirer les données avec remise, conduisant à une loi échangeable. L'espace des labels $\mathcal{Y}$ est supposé fini (e.g. $\mathcal{Y} = \{-1, +1\}$).

On considère un/des modèles de classement :

$$\{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\},$$

et on cherche à minimiser une fonction de perte moyenne

$$\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} \ell\left[f_\theta(X_i), Y_i\right]\right\} \stackrel{\text{def}}{=} \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} W_i(\theta)\right),$$

par exemple $\ell(y', y) = \mathbb{1}(y' \neq y)$ (erreur de classification). On n'est pas loin du problème générique de régression :

$$\inf_\theta \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} W_i(\theta)\right),$$

si ce n'est que l'on peut exploiter parfois le fait que les $W_i(\theta)$ sont à valeurs dans un ensemble fini (e.g. $\{0, 1\}$).

Minimisation du risque

Étant donnée une fonction de risque

$$R(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} W_i(\theta)\right],$$

et le risque empirique dont elle est l'espérance

$$r(\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i(\theta),$$

comment faire pour minimiser $R(\theta)$ en observant uniquement $r(\theta)$?

Un intervalle de confiance pour l'espérance d'une variable aléatoire dont on ne sait pas grand-chose, premier essai, la transformée de Laplace :

$$\mathbb{E}\exp\big[-\lambda r(\theta)\big] = \prod_{i=1}^{N} \mathbb{E}\Big\{\exp\big[-\tfrac{\lambda}{N} W_i(\theta)\big]\Big\}$$

$$= \prod_{i=1}^{N}\Big\{1 - \frac{\lambda}{N}\mathbb{E}\big[W_i(\theta)\big] + \frac{\lambda^2}{2N^2}\mathbb{E}\big[W_i(\theta)^2\big] + \cdots\Big\},$$

qui vous conduit à de graves ennuis (hypothèses sur les moments de $W_i(\theta)$ ...).

Deuxième tentative, la transformée de Laplace, mais d'un risque empirique modifié :

$$\mathbb{E}\exp\left\{\sum_{i=1}^{N}\log\left[1-\frac{\lambda}{N}\left[W_i(\theta)\wedge\tfrac{N}{\lambda}\right]\right]\right\}$$

$$=\prod_{i=1}^{N}\left\{1-\frac{\lambda}{N}\mathbb{E}\left[W_i(\theta)\wedge\tfrac{N}{\lambda}\right]\right\}$$

$$\leq\left\{1-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[W_i(\theta)\wedge\tfrac{N}{\lambda}\right]\right\}^{N}.$$

D'où l'idée de poser

$$\tilde{r}_\lambda(\theta) = -\frac{1}{\lambda}\sum_{i=1}^{N}\log\left\{1-\frac{\lambda}{N}\left[W_i(\theta)\wedge\tfrac{N}{\lambda}\right]\right\},$$

$$= r(\theta)+\frac{\lambda}{2N}m_\lambda(\theta),$$

$$\tilde{R}_\lambda(\theta) = \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[W_i(\theta)\wedge\tfrac{N}{\lambda}\right]\right\} \simeq R(\theta) \text{ (quand } \tfrac{N}{\lambda} \text{ est grand)},$$

pour obtenir

$$\mathbb{E}\exp\left\{\lambda\left[\tilde{R}_\lambda(\theta)-\tilde{r}_\lambda(\theta)\right]\right\}$$

$$\leq \mathbb{E}\exp\left\{\lambda\left\{-\tfrac{N}{\lambda}\log\left[1-\tfrac{\lambda}{N}\tilde{R}_\lambda(\theta)\right]-\tilde{r}_\lambda(\theta)\right\}\right\} \leq 1,$$

et donc pour θ fixé, avec probabilité $1-\epsilon$

$$\tilde{R}_\lambda(\theta) \leq r_\lambda(\theta)+\frac{\lambda}{2N}m_\lambda(\theta)-\frac{\log(\epsilon)}{\lambda}.$$

Uniformité en θ ? on va mesurer l'espace des paramètres Θ avec une probabilité a priori $\pi \in \mathcal{M}_+^1(\Theta)$, et réintégrer la transformée de Laplace à l'intérieur de l'espérance en utilisant Fubini

$$\mathbb{E}\left\{ \int_\Theta \pi(d\theta) \exp\left[\lambda \left[\tilde{R}_\lambda(\theta) - \tilde{r}_\lambda(\theta) \right] \right] \right\} \leq 1.$$

On peut alors traiter un choix de θ dépendant des données décrit par une loi a posteriori $\rho(d\theta \mid W) \ll \pi(d\theta)$ en écrivant simplement

$$\mathbb{E}\left\{ \exp\left\{ \lambda \int \rho(d\theta \mid W) \left[\tilde{R}_\lambda(\theta) - \tilde{r}_\lambda(\theta) \right] - \mathcal{K}\left[\rho(\cdot \mid W), \pi \right] \right\} \right\}$$

$$\leq \mathbb{E}\left\{ \int_\Theta \rho(d\theta \mid W) \exp\left[\lambda \left[\tilde{R}_\lambda(\theta) - \tilde{r}_\lambda(\theta) \right] - \log\left(\frac{d\rho(\cdot \mid W)}{d\pi} \right) \right] \right\} \leq 1.$$

D'où avec probabilité $1-\epsilon$ sous la loi jointe

$$(W,\theta) \sim \left(\otimes_{i=1}^{N} P_i\right)(dw)\cdot\rho(d\theta|w),$$

$$\tilde{R}_\lambda(\theta) \leq \tilde{r}_\lambda(\theta) + \frac{1}{\lambda}\left[\log\left(\frac{d\rho}{d\pi}\right) - \log(\epsilon)\right],$$

et avec probabilité $1-\epsilon$ sous la loi de l'échantillon W

$$\int \rho(d\theta|W)\tilde{R}_\lambda(\theta) \leq \int \rho(d\theta|W)\tilde{r}_\lambda(\theta) + \frac{1}{\lambda}\left[\mathcal{K}(\rho(.|W),\pi) - \log(\epsilon)\right].$$

π optimale pour ρ fixée : $\mathbb{E}\left[\rho(.|W)\right]$.
ρ optimale pour π fixée : $\pi_{\exp(-\lambda\tilde{r}_\lambda)}$ définie par sa densité suivant la formule générale $\frac{d\pi_h}{d\pi} = \frac{h}{\int \pi(d\theta)h(\theta)}$.

L'information mutuelle entre paramètre et échantillon contrôle l'erreur d'estimation due à la taille du modèle : pour ρ fixée, et $\pi = \mathbb{E}\big[\rho(.|W)\big]$ (malheureusement pas observable !), $\mathbb{E}\big[\mathcal{K}(\rho(.|W), \pi)\big]$ est égale à l'information mutuelle entre l'échantillon W et le paramètre θ sous la loi jointe $\left(\otimes_{i=1}^{N} P_i\right) \cdot \rho(.|w)$ (c'est à dire le nombre de bits que l'on peut économiser dans la transmission de θ après avoir transmis W : en particulier pour des v.a. indépendantes, l'information mutuelle est nulle).
Une technique pour obtenir des garanties sur le comportement d'un estimateur $\hat{\theta}$ consiste donc à tronquer la précision de $\hat{\theta}$, soit en choisissant $\hat{\theta}$ au hasard dans un voisinage du minimiseur de $r(\theta)$, soit en minimisant $r(\theta)$ sur un réseau, de façon à diminuer l'information mutuelle entre $\hat{\theta}$ et l'échantillon.

Vicissitudes et/ou améliorations techniques :

- la meilleure vitesse d'estimation de $\arg\min_\theta R(\theta)$ ne dépend pas des variances des $r(\theta)$, mais des variances des différences $r(\theta) - r(\theta')$, il faut donc tout refaire pour la famille à deux paramètres de variables aléatoires

$$W_i'(\theta, \theta') = W_i(\theta) - W_i(\theta'), \qquad (\theta, \theta') \in \Theta^2.$$

- Pour atteindre les vitesses d'estimation minimax, il faut « mesurer »l'espace des paramètres Θ avec des lois a priori π suffisamment proche de $\mathbb{E}\big[\rho(\cdot|W)\big]$, par exemple des lois localisées $\pi_{\exp(-\beta R)}$. Se pose alors le problème de l'estimation de $\mathcal{K}\big[\rho(.|W), \pi_{\exp(-\beta R)}\big]$ ou de $\log\left[\frac{d\rho(\cdot|W)}{d\pi_{\exp(-\beta R)}}\right]$, suivant le type de bornes utilisées.

minimisation du risque

Les deux marchent ensemble, suivant les formules *(ici dans le cas borné)*

$$\int \nu(d\theta')\big[R(\theta) - R(\theta')\big] \leq \int \nu(d\theta')\tilde{r}'_\lambda(\theta,\theta') + \frac{1}{\lambda}\log\left[\epsilon^{-1}\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\theta)\right]$$

avec $\left(\otimes_{i=1}^N P_i\right)\cdot\rho(\cdot|w)$ probabilité $1-\epsilon$,

$$\log\left[\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\theta)\right] = \beta R(\theta) + \log\left[\frac{d\rho}{d\pi}(\theta)\right] + \log\left\{\int \pi(d\theta')\exp\big[-\beta R(\theta')\big]\right\}$$

$$= \sup_\nu \beta\int \nu(d\theta')\big[R(\theta) - R(\theta')\big] + \log\left[\frac{d\rho}{d\pi}(\theta)\right] - \mathcal{K}(\nu,\pi)$$

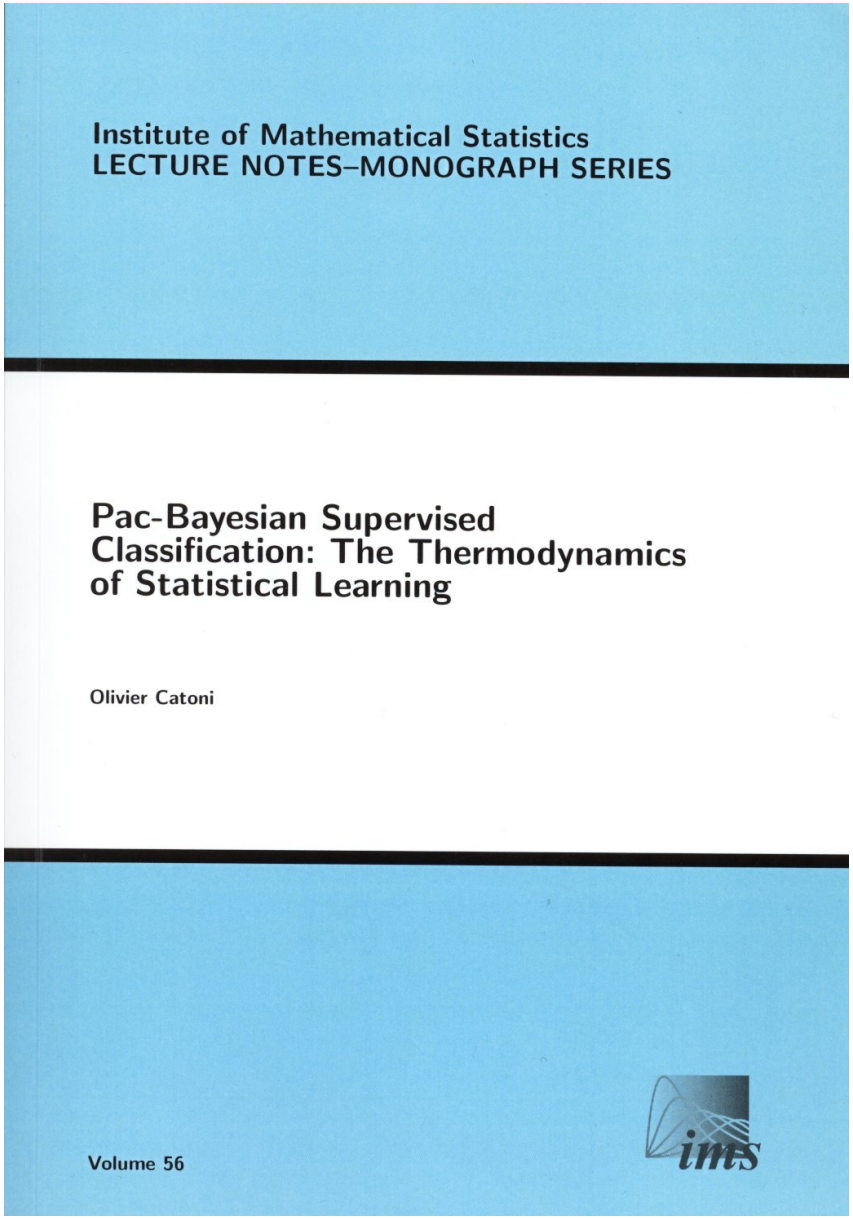
$$\leq \sup_\nu \beta\int \nu(d\theta')\tilde{r}'_\lambda(\theta,\theta') + \frac{\beta}{\lambda}\log\left[\epsilon^{-1}\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\theta)\right] + \log\left[\frac{d\rho}{d\pi}(\theta)\right] - \mathcal{K}(\nu,\pi),$$

avec probabilité $1-\epsilon$.

L'entropie se prête à toutes sortes de manipulations algébriques liées au fait qu'elle est la transformée de Legendre de la transformée de Laplace

$$h \mapsto \log\Big\{ \int \pi(d\theta) \exp\big[h(\theta)\big] \Big\}$$
$$= \sup_{\nu \in \mathcal{M}_+^1(\Theta)} \int \nu(d\theta) h(\theta) - \mathcal{K}(\nu, \pi)$$
$$= \int \nu(d\theta) h(\theta) - \mathcal{K}(\nu, \pi) + \mathcal{K}(\nu, \pi_{\exp(h)}).$$

On arrive ainsi à fabriquer des méthodes de régression et de sélection de modèles pour lesquelles on peut calculer un intervalle de confiance (à droite) sous des hypothèses très faibles, qui se trouve être d'ordre optimal dans de nombreux cas particuliers, paramétriques ou non paramétriques.



Pour continuer l'aventure, ce titre téléchargeable sur ArXiv, ainsi que les travaux de Jean-Yves Audibert et Pierre Alquier !

Classification non supervisée et codage

Dans la discussion précédente, on a vu apparaître une probabilité a priori π, qui ne servait pas à décrire un phénomène aléatoire, mais à borner l'information mutuelle entre paramètre et échantillon. Cette loi a priori définit un certain codage implicite du paramètre. En effet (tout au moins dans le cas où Θ est fini) tout code binaire préfixe $c : \Theta \to \{0,1\}^*$ définit une sous-probabilité en vertu de l'inégalité de Kraft

$$\sum_{\theta} 2^{-\ell[c(\theta)]} \leq 1,$$

où $\ell\big[c(\theta)\big]$ désigne la longueur (nombre de bits) de $c(\theta)$. Réciproquement, à partir de toute (sous)-probabilité π sur Θ, on peut définir (au moins) un code binaire préfixe tel que $\big|\ell[c(\theta)] + \log_2\big[\pi(\theta)\big]\big| \leq 1$.

- Les techniques de classification supervisée permettent de choisir une classification parmi un ou plusieurs modèles de classement en maîtrisant la question du surapprentissage. Elle ne permettent pas de trouver un bon modèle de classement.
- La pertinence des modèles génériques de classement (arbres de décision, séparation par des hyperplans, ...) dépend de la représentation des données à laquelle ils sont appliqués.
- Inversement, lorsque l'on connaît la loi conditionnelle des classes, la classification optimale s'obtient en faisant le rapport des vraisemblances, donc en comparant les longueurs des codes associés.

Une manière d'aborder le problème de la représentation consiste donc à estimer à partir des données non étiquetées $(X_1, \ldots, X_N)$ une probabilité qui minimise la longueur moyenne du code associé, c'est-à-dire à faire de l'estimation de densité en maximisant la log vraisemblance. Dans le contexte de la classification, il est naturel de considérer des modèles de mélange où $\frac{dP}{dQ} = \int \pi(d\theta) f_\theta(x)$. On peut alors chercher à maximiser $\mathbb{E}\left\{\log\left[\int \pi(d\theta) f_\theta(x)\right]\right\}$. On en tire une loi jointe P_{jointe} sur le couple (θ, x), et on peut représenter $x \in \mathcal{X}$ par $P_{\text{jointe}}(\theta \mid x)$. Ainsi, on peut associer à un modèle de classification de θ un modèle de classification de x : le modèle de mélange $\pi(d\theta) f_\theta(x)$ définit une dualité entre $\mathcal{X}$ et Θ qui permet de représenter le problème de classification initial dans Θ.

Application à la classification d'images

(en cours)

Question : classer des images en fonction de leur contenu (e.g. retrouver sur internet les photos des Invalides).
De nouvelles difficultés (classiques !) se présentent :

- Où chercher l'objet (problème de segmentation) ?
- Comment prendre en compte les variations dues
  - aux changements d'éclairage ?
  - aux changements de point de vue (translations, rotations, plus généralement transformations projectives) ?

  (prise en compte des invariants)

Réponse possible aux variations d'éclairage : détection de contours.
Classification locale multiéchelle des pixels :

- plus lumineux que l'environnement ;
- aussi lumineux que l'environnement ;
- moins lumineux que l'environnement ;

On définit les contours comme étant les bords de ces classes.

On prend comme espace d'état $\mathcal{X} = \{0,1\}^I$, où I est une grille plus fine que celles des pixels. On représente les niveaux de gris par des textures sur $\{0,1\}$ à un niveau de discrétisation plus fin que le niveau observé.
Une image est donc représentée au départ par $(Z_i)_{i\in I} \in \{0,1\}^I$.
On réplique cette représentation en $d+1$ copies (identiques) $Z^0, \ldots, Z^d$.

On code les répliques à l'aide d'une loi jointe telle que

$$P(Z^k|Z^n, n<k) = \prod_{i\in I} P(Z_i^k|Z_j^0, j\in i+V_k), \quad 0<k\leq d,$$

où V_k est un voisinage de 0 et où $P(Z_i^k|Z_j^0, j\in i+V_k)$ est la Bernoulli de paramètre $p_i^k = \frac{1}{|V_k|}\sum_{j\in i+V_k} Z_j^0$. On considère alors les lois Q_+^k et Q_-^k, définies par les mêmes équations que P à ceci prêt que $Q_+^k(Z_i^k|Z^0)$ est la Bernoulli de paramètre $\left(p_i^k + \frac{\beta}{|V_k|}\right)\wedge 1$ et $Q_-^k(Z_i^k|Z^0)$ est la Bernoulli de paramètre $\left(p_i^k - \frac{\beta}{|V_k|}\right)\vee 0$.

On fusionne les échelles en considérant

$$Q_+(Z) = \frac{1}{d}\sum_{k=1}^{d} Q_+^k(Z) = Q_+(Z^0)\prod_{i\in I} Q_+\left(Z_i^k, 0 < k \leq d\right).$$

On établit alors sur chaque imagette correspondant à chaque pixel deux classements binaires, en regardant les densités de Q_+ et Q_- par rapport à la loi de référence P. On obtient les contours en prenant les bords de ces deux images seuillées à 1 (ou en répétant l'opération).

Pour avoir les orientations des bords, on peut alors calculer les gradients de Q_+ et Q_-, à l'aide d'un opérateur de convolution, du type

$$G_i = \sum_{j \in I} \rho(i-j) Z_j(i-j) \in \mathbb{R}^2,$$

où ρ est un noyau de convolution. On peut le restreindre aux contours en formant $C_i = B_i G_i$, où B_i est l'indicatrice des contours.
On obtient ainsi des contours représentés par un point, une intensité et une direction (normale).

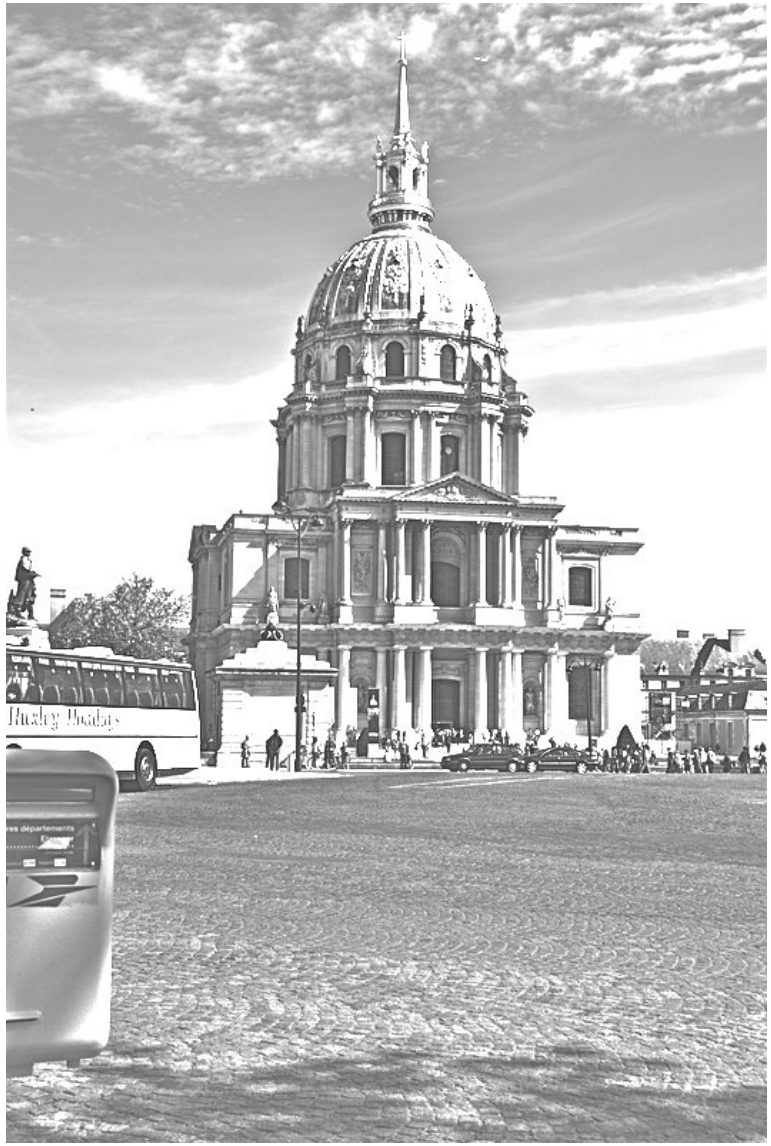
On peut alors construire une représentation duale multiéchelle des contours en formant le champ de vecteur

$$G'^{k}_{i} = \sum_{j \in I} 2^{-2k} \rho\left[2^{-k}(i-j)\right] \langle i-j, C_j \rangle C_j.$$

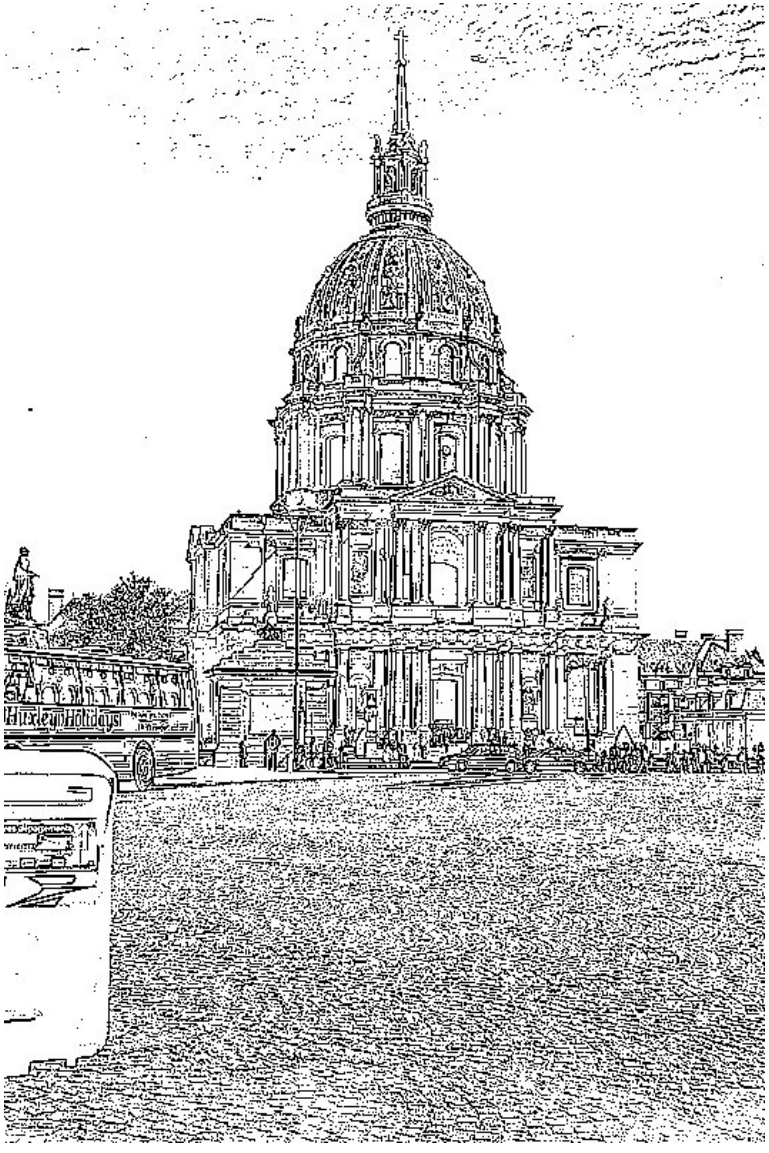
Ces champs possèdent la propriété intéressante d'avoir des zéros isolés qui correspondent à des angles ou à des centres de symétrie. On peut donc espérer les utiliser pour résoudre le problème de la mise en correspondance de plusieurs vues (et par la même occasion offrir une approche de la question de la segmentation). Les intersections peuvent servir à former des repères projectifs (il en faut 4), ce qui constitue une piste vers l'invariance projective de la classification (le fait de pouvoir reconnaître des scènes planes prises sous des angles différents paraissant souhaitable).

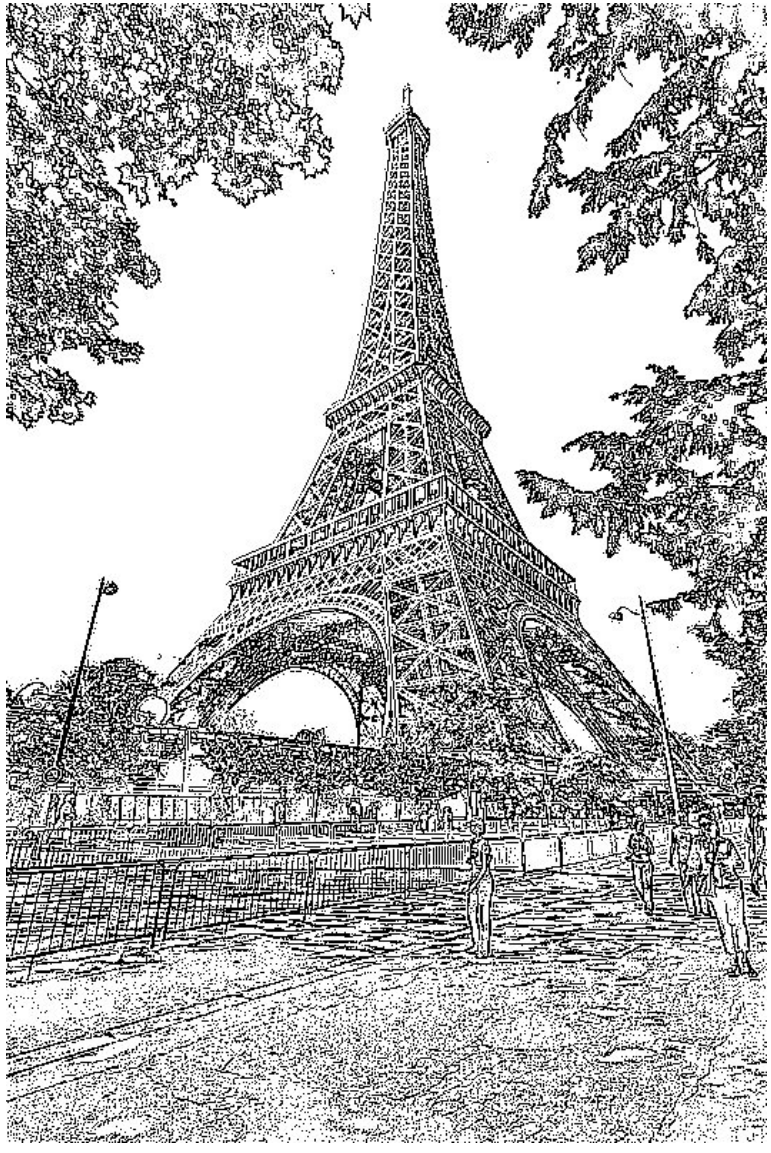
classification d'images

classification d'images

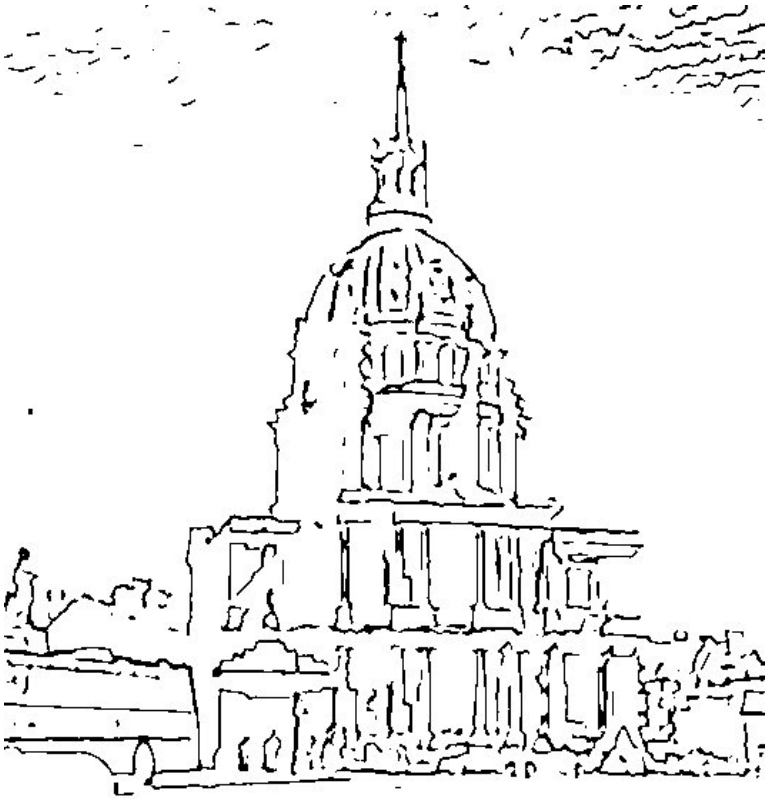


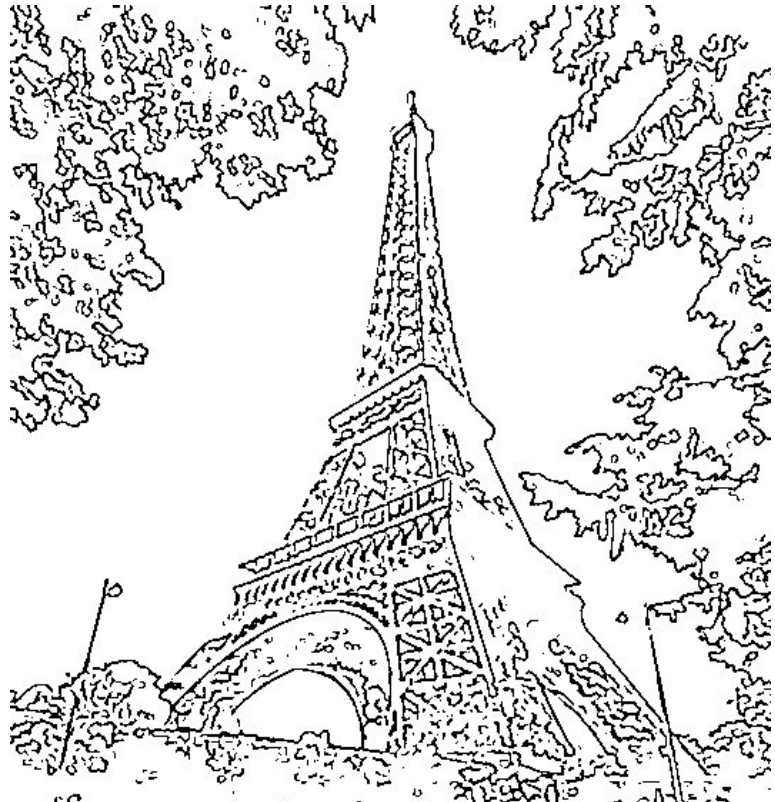
classification d'images





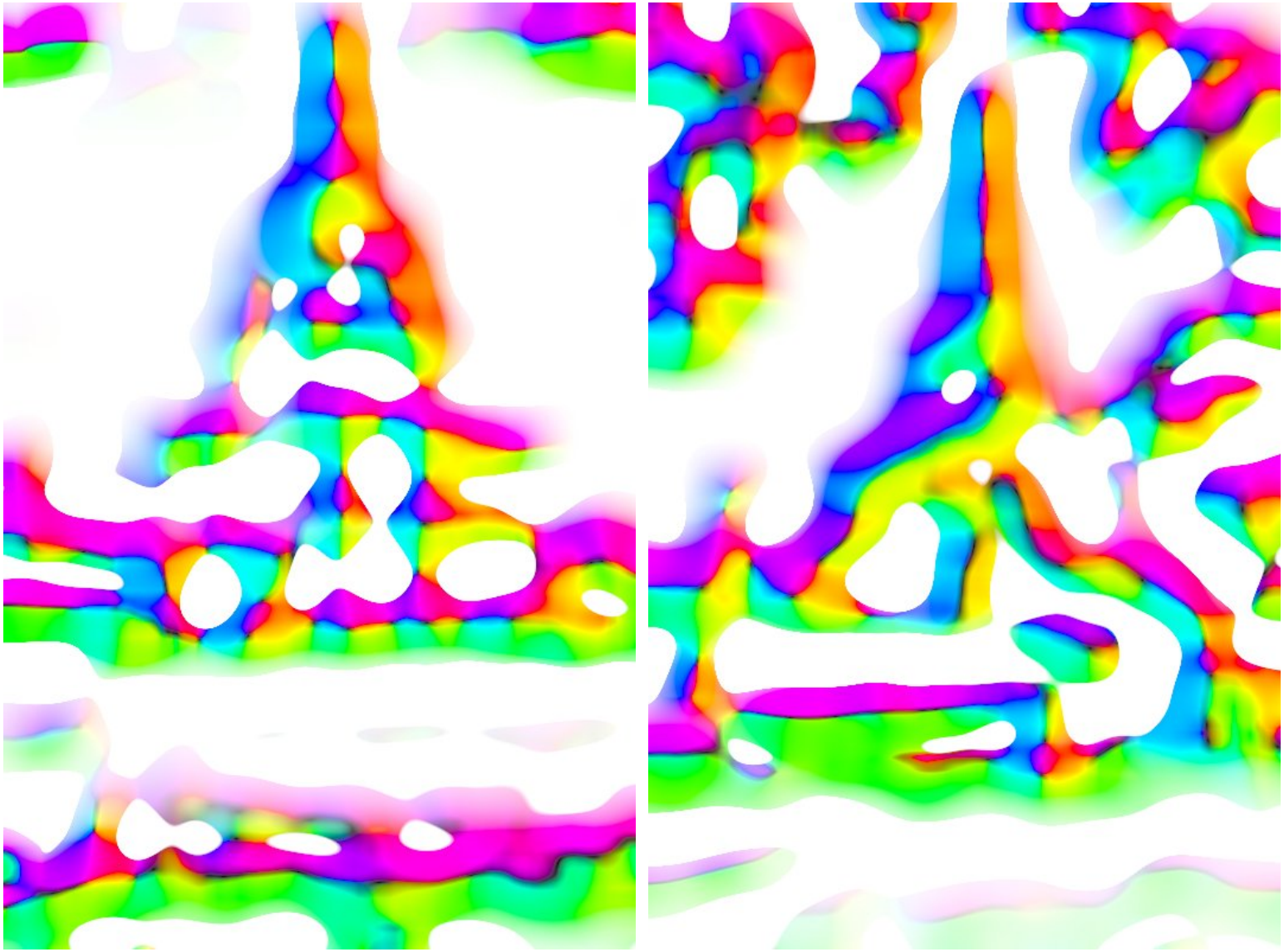
classification d'images





classification d'images

classification d'images



classification d'images