Information k means and application to digital images

Gautier Appert

ENSAE Paris Tech.
joint work with Olivier Catoni

Representation of image with multi-level bags of labels

First step: creating a bag of labels

- Divide each image $\{X_1, \ldots, X_n\}$ into non overlapping patches $\{B_i, i \in I\}$ $\Longrightarrow$ cluster all patches $\{B_j, \text{for all } j\}$.

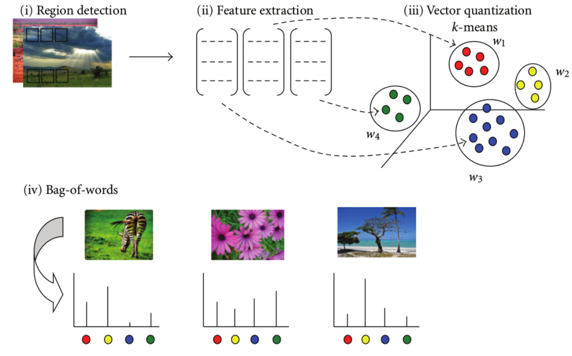

Figure : Firgure 1 in Bag-of-Words Representation in Image Annotation: A Review. Chih-Fong Tsai

- Each image X is then represented as $\mathbb{P}_{W|X} = \frac{1}{m} \sum_{i=1}^{m} \delta_{w_i}$.

Representation of image with multi-level bags of labels

Multi-level

- Clustering labels with respect to a distortion measure is difficult $\Longrightarrow$ Instead use contextual modeling : cluster labels w and w' if they share the same context, but do not appear together. Define the context C of W in image X as $C = \mathbb{P}_{W|X} \circ f_{W,\Delta}^{-1}$, where $f_{W,\Delta}$ is the function that sends W to the outer state Δ. Note that $C \in \mathcal{M}_+^1\big(\mathcal{W} \cup \{\Delta\}\big)$ is a random measure, and a function of the couple of random variables (X, W).
- Cluster/agreggate words $\{w, w'\}$ **iff** $\mathbb{P}_{C|W=w} \simeq \mathbb{P}_{C|W=w'}$.
- Which means that we are looking for a classification function $\ell : \mathcal{W} \to \mathcal{Z}$ such that
$$\mathbb{P}_{C|W} = \mathbb{P}_{C|\ell(W)} \iff C \perp\!\!\!\perp W \mid \ell(W).$$
- When this is the case, $\mathbb{P}_{\mathbb{P}_{W|X}}$ can be recovered from $\mathbb{P}_{\mathbb{P}_{\ell(W)|X}}$ and $\mathbb{P}_{W|\ell(W)}$.

Euclidian k-means: theoretical and empirical loss

Let $X \in \mathbb{R}^d$ with $\mathbb{P}_X\left[\|X\|_2^2\right] < \infty$ and let $\ell : \mathcal{X} \to \{1, \ldots, k\}$ be the labelling function. The set $\mathcal{X}$ can be $\mathbb{R}^d$ or the index set $\{1, \ldots, n\}$.

theoretical loss

$$\inf_{\ell} \inf_{\mu_1,\ldots,\mu_k} \mathbb{P}_X\left[\|X - \mu_{\ell(X)}\|_2^2\right]$$
$$= \inf_{\ell} \mathbb{P}_X\left[\|X - \mathbb{E}[X|\ell(X)]\|_2^2\right]$$
$$= \inf_{\mu_1,\ldots,\mu_k} \mathbb{P}_X\left[\min_{j \le k} \|X - \mu_j\|_2^2\right]$$

Empirical loss

$$\inf_{\ell} \inf_{\mu_1,\ldots,\mu_k} \frac{1}{n}\sum_{i=1}^{n} \|X_i - \mu_{\ell(i)}\|_2^2$$
$$= \inf_{\ell} \frac{1}{n}\sum_{j=1}^{k} \sum_{i \in \ell^{-1}(j)} \|X_i - \overline{\mu}_j\|^2$$
$$= \inf_{\mu_1,\ldots,\mu_k} \frac{1}{n}\sum_{i=1}^{n} \min_{j \le k} \|X_i - \mu_j\|_2^2$$

Lloyd 's algorithm finds a local minimum through an iterative scheme: allocate data points to the nearest centroid and recompute centers from this partition.

Geometric mean of a conditional probability measure

Definition (Geometric mean of a conditional probability measure.)

Define the geometric mean function $\mathcal{G}(\cdot,\cdot)$ of a conditional probability measure $\mathrm{d}P(t|s) = m(t|s)\,\mathrm{d}\mu(t)$ with respect to the probability measure $\mathrm{d}P(s)$ as

$$\mathcal{G}\Big(\mathrm{d}P(t|s), \mathrm{d}P(s)\Big) = Z^{-1}\ \exp\left\{\int \log\Big[\mathrm{d}P(t|s)\Big]\ \mathrm{d}P(s)\right\}$$
$$\stackrel{\text{def}}{=} Z^{-1}\ \exp\left\{\int \log\Big[m(t|s)\Big]\ \mathrm{d}P(s)\right\}\ \mathrm{d}\mu(t),$$

where Z is a normalizing constant. Note that this is independent from the choice of μ: if $\nu \in \mathcal{M}_+^1$ is such that $\mu \ll \nu$, $\mathrm{d}P(t|s) = \dfrac{\mathrm{d}\mu}{\mathrm{d}\nu}(t)\, m(t|s)\, \mathrm{d}\nu(t)$ and
$$\mathcal{G}\Big(\mathrm{d}P(t|s), \mathrm{d}P(s)\Big) = Z^{-1}\ \exp\left\{\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(t)\ m(t|s)\right)\ \mathrm{d}P(s)\right\}\ \mathrm{d}\nu(t).$$

Information k-means: theoretical and empirical loss

Let (Y, X) be a couple of random variables, assume that $\mathbb{P}_{Y|X}$ is known, whereas $\mathbb{P}_X$ may be unknown.

Theoretical version

$$\inf_{\ell} \inf_{Q_{Y|\ell(X)}} \mathbb{P}_X \Big[\mathcal{K}\big(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}\big) \Big] = \inf_{Q_{Y|j},\, j \leq k} \mathbb{P}_X \Big[\inf_{j \leq k} \mathcal{K}\big(Q_{Y|j}, \mathbb{P}_{Y|X}\big) \Big]$$

$$= \inf_{\ell} \mathbb{P}_X \Big[\mathcal{K}\big(Q^*_{Y|\ell(X)}, \mathbb{P}_{Y|X}\big) \Big] = \inf_{\ell} \mathbb{P}_X \Big[\log\big(Z^{-1}_{\ell(X)}\big) \Big],$$

where $Q^*_{Y|\ell(X)}$ (the information k-means centers) and $Z_{\ell(X)}$ (the normalizing constants) are defined as

$$\mathbb{Q}^*_{Y|\ell(X)} \overset{\text{def}}{=} \mathcal{G}\big(\mathbb{P}_{Y|X}, \mathbb{P}_{X|\ell(X)}\big) = Z^{-1}_{\ell(X)} \exp\Big\{ \mathbb{P}_{X|\ell(X)} \Big[\log \mathbb{P}_{Y|X} \Big] \Big\}.$$

Information k-means: theoretical and empirical loss

Consider a set of conditional probability distributions $R_{Y|i}$ for the random variable Y knowing $i \in [n]$, that we want to cluster.

Empirical loss

$$\begin{aligned}
&\inf_{\ell:\{1,\dots,n\}\rightarrow\{1,\dots,k\}} \inf_{Q_{Y|\ell(i)}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}\Big(Q_{Y|\ell(i)}, R_{Y|i}\Big) \\
&\quad = \inf_{Q_{Y|j}} \frac{1}{n}\sum_{i=1}^{n} \inf_{j\in\{1,\dots,k\}} \mathcal{K}\Big(Q_{Y|j}, R_{Y|i}\Big) \\
&\quad = \inf_{\ell} \frac{1}{n}\sum_{j=1}^{k} \inf_{Q_{Y|j}} \sum_{i\in\ell^{-1}(j)} \mathcal{K}\Big(Q_{Y|j}, R_{Y|i}\Big) \\
&\quad = \inf_{\ell} \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}\Big(Q^*_{Y|\ell(i)}, R_{Y|i}\Big) = \inf_{\ell} \sum_{j=1}^{k} \frac{\big|\ell^{-1}(j)\big|}{n} \log\Big(Z_j^{-1}\Big),
\end{aligned}$$

where $Q^*_{Y|j} = Z_j^{-1} \prod_{i\in\ell^{-1}(j)} R_{Y|i}^{1/|\ell^{-1}(j)|}$.

Information k-means: theoretical and empirical loss

Starting point

Due to the properties of the Kullback divergence, the following algorithm to compute an initial classification ℓ gives promising results.

- Start from $k = 1$ and $\ell^{-1}(1) = \{1, \dots, n\}$.
- Switch from k to $k+1$ by removing iteratively from $\ell^{-1}(k)$ $\arg\max_{i \in \ell^{-1}(k)} \mathcal{K}\big(Q^*_{Y|k}, R_{Y|i}\big)$ to put it in $\ell^{-1}(k+1)$, until $\log\big(Z_k^{-1}\big) \leq \eta$.
- Continue if $\log\big(Z_{k+1}^{-1}\big) > \eta$.

Link with Information projection

Definition (Information projection.)

Let P be a probability distribution, and let $\mathcal{Q}$ be set of probability distribution. The information projection or I-projection of P onto $\mathcal{Q}$ is defined as

$$Q^* \in \arg\min_{Q \in \mathcal{Q}} \mathcal{K}\big(Q, P\big).$$

Information k-means seen as an information projection

- Consider the model

$$\mathcal{Q} = \Big\{ \mathbb{Q}_{Y,X} : \mathbb{Q}_X = \mathbb{P}_X,\ \mathbb{Q}_{Y|X} = \mathbb{Q}_{Y|\ell(X)},\ \ell(X) \in \{1, \dots, k\} \Big\},$$

$$\begin{aligned} &\inf_{\mathbb{Q}_{Y,X} \in \mathcal{Q}} \mathcal{K}\big(\mathbb{Q}_{Y,X}, \mathbb{P}_{Y,X}\big) \\ &\quad = \inf_{\mathbb{Q}_{Y,X} \in \mathcal{Q}} \mathbb{Q}_X \Big[\mathcal{K}\Big(\mathbb{Q}_{Y|X}, \mathbb{P}_{Y|X} \Big) \Big] + \mathcal{K}\Big(\mathbb{Q}_X, \mathbb{P}_X \Big) \\ &\quad = \inf_{\ell, \mathbb{Q}_{Y|\ell(X)}} \mathbb{P}_X \Big[\mathcal{K}\Big(\mathbb{Q}_{Y|\ell(X)}, \mathbb{P}_{Y|X} \Big) \Big]. \end{aligned}$$

Information k-means become Euclidian with Gaussian distribution

Information k-means generalize Euclidian k-means

- Take $\mathbb{P}_{Y|X} = \mathcal{N}_p\big(X, \Sigma\big)$.
- One obtains

$$
\begin{aligned}
\mathrm{d}Q^*_{Y|\ell(X)}(y) &\propto \exp\left\{\mathbb{P}_{X|\ell(X)}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\lambda}(y)\right)\right]\right\}\,\mathrm{d}\lambda(y) \\
&\propto \exp\left\{-\frac{1}{2}\left(y^\top\Sigma^{-1}y - 2y^\top\Sigma^{-1}\mathbb{E}\big[X|\ell(X)\big]\right)\right\} \\
&\propto \mathcal{N}_p\Big(\mathbb{E}\big[X|\ell(X)\big], \Sigma\Big)
\end{aligned}
$$

- Then $\mathcal{K}\Big(Q^*_{Y|\ell(X)}, \mathbb{P}_{Y|X}\Big) = \Big\|X - \mathbb{E}\big[X|\ell(X)\big]\Big\|^2_{\Sigma^{-1}}$.
- $\inf_{\ell}\ \mathbb{P}_X\Big[\mathcal{K}\Big(Q^*_{Y|\ell(X)}, \mathbb{P}_{Y|X}\Big)\Big] = \inf_{\ell}\ \mathbb{P}_X\Big[\Big\|X - \mathbb{E}\big[X|\ell(X)\big]\Big\|^2_{\Sigma^{-1}}\Big]$

Information k-means loss in the case of discrete $\mathbb{P}_Y$

Let $Y \in \mathcal{Y}$ with $|\mathcal{Y}| < \infty$.

- $\mathcal{L}(Q) = \mathbb{P}_X\Big[\min_{i \leq k} \mathcal{K}\big(Q_{Y|i}, \mathbb{P}_{Y|X}\big)\Big]$
- Put $q_i = \frac{\mathrm{d}Q_{Y|i}}{\mathrm{d}\nu}$, $p_X = \frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\nu} \implies \mathcal{L}(q) = \mathbb{P}_X\Big[\min_{i \leq k} \mathcal{K}\big(q_i, p_X\big)\Big]$.

- Recall $\mathcal{K}\big(q_i, p_X\big) = \langle q_i, \log(q_i) - \log(p_X)\rangle$.
- Put $\theta_i = (-q_i, \langle q_i, \log(q_i)\rangle)^\top$,
  $\theta_{i,j} = (q_i - q_j, \langle q_i, \log(q_i)\rangle - \langle q_i, \log(q_i)\rangle)^\top \in \mathbb{R}^{|\mathcal{Y}|+1}$ and
  $W = (\log p_X, 1)^\top \in \mathbb{R}^{|\mathcal{Y}|+1}$.
- Hence, $\mathcal{K}\big(q_i, p_X\big) = \langle \theta_i, W\rangle$ and
  $\mathcal{K}\big(q_i, p_X\big) < \mathcal{K}\big(q_j, p_X\big) \iff \langle \theta_{i,j}, W\rangle \geq 0$.

upper bound on the Loss

- Using the fact that

$$\min_{i \leq k} a_i = \sum_{i=1}^{k} a_i \prod_{j=1}^{i-1} \mathbb{1}\left(a_i < a_j\right) \prod_{j=i+1}^{k} \mathbb{1}\left(a_i \leq a_j\right).$$

- We can rewritte

$$\mathcal{L}(q) \leq \sum_{i=1}^{k} \mathbb{P}_X \left[\langle \theta_i, W \rangle \prod_{j \neq i} \mathbb{1}\left(\langle \theta_{i,j}, W \rangle \geq 0\right) \right].$$

PAC-Bayesian Margin bounds on information k-means

Put a perturbation and a margin

- Gaussian perturbation $\rho_\theta = \mathcal{N}\left\{\theta, \beta^{-1}\mathrm{I}_{|\mathcal{Y}|+1}\right\}$.
- Margin $M = \gamma\|W\|$.

lemma

$$\mathcal{L}(q) \leq \sum_{i=1}^{k} \Phi\left(\gamma\sqrt{\beta}\right)^{-(k-1)} \times \mathbb{P}_X\left[\langle\theta_i, W\rangle \prod_{j\neq i} \int \mathbb{1}\left(\langle\theta_{i,j}^{'}, W\rangle + \gamma\|W\| \geq 0\right) \mathrm{d}\rho_{\theta_{i,j}}(\theta_{i,j}')\right]$$

- Looks like some kind of classification problem with margin $M = \gamma\|W\|$. [Catoni, Lecture notes , 2014].
- Estimation of the mean of $\langle\theta_i, W\rangle$. [Catoni,Giulini 2017].

PAC-Bayesian Margin bounds on information k-means

Upper bound of the information k-means loss

Introduce $g_1(t) = \frac{1}{t}(\exp(t) - 1)$ and $g_2 = \frac{1}{t^2}(\exp(t) - 1 - t)$. With probability at least $1 - \varepsilon$,

$$\mathcal{L}(q) \leq \Phi\left(\gamma\sqrt{\beta}\right)^{-(k-1)} \left\{ \sum_{i=1}^{k} \widehat{\mathbb{P}}_X^n \left[\langle \theta_i, Z \rangle \overline{H}(W, \theta_{-i}) \right] + \frac{\lambda a}{2} \, \mathbb{P}_X\left(\langle \theta_i, W \rangle^2 \overline{H} \right) + \frac{\lambda b}{2\beta} \, \mathbb{P}_X\left(\|W\|^2 \overline{H} \right) + \frac{\alpha^p}{p+1} \, \mathbb{P}_X\left(\left| \langle \theta_i, W \rangle \right| \, \|W\|^p \right) + \frac{k \, \beta}{2n\lambda} \sum_{i=1}^{k} \left\{ \|\theta_i\|^2 + \sum_{j \neq i} \|\theta_{ij}\|^2 \right\} \frac{k \ \log(\varepsilon^{-1})}{n\lambda} \right\},$$

$$\text{where } a = g_2\left(\frac{\lambda \|\theta_i\|}{\alpha} \right), \quad b = g_1\left(\frac{\lambda^2}{2\beta\alpha^2} \right) \exp\left(\frac{\lambda \|\theta_i\|}{\alpha} \right),$$

$$\overline{H} = \prod_{j \neq i} \Phi\left(\sqrt{\beta}\left(\gamma + \|W\|^{-1} \langle \theta_{i,j}, W \rangle\right) \right) \text{ and } Z = \frac{\min(\lambda\|W\|, 1))W}{\lambda\|W\|}.$$

PAC-Bayesian Margin bounds on Euclidian k-means

Euclidian $k - means$ loss

- Consider $\mathcal{L}(\mu) = \mathbb{P}_X\left(\min_{i \le k} \|X - \mu_i\|^2\right)$

Change of notation

- Put $\theta_i = (-\mu_i, \|\mu_i\|^2)^\top$, $\theta_{i,j} = (\mu_i - \mu_j, \|\mu_j\|^2 - \|\mu_i\|^2)^\top \in \mathbb{R}^{p+1}$ and $W = (2X, 1)^\top \in \mathbb{R}^{p+1}$.
- Hence (polarization identity), $\|X - \mu_i\|^2 = \langle \theta_i, W \rangle + \|X\|^2$ and $\|X - \mu_i\|^2 < \|X - \mu_j\|^2 \iff \langle \theta_{i,j}, W \rangle \ge 0$.

$$\mathcal{L}(\mu) \le \sum_{i=1}^{k} \mathbb{P}_X\left[\langle \theta_i, W \rangle \prod_{j \ne i} \mathbb{1}\left(\langle \theta_{i,j}, W \rangle \ge 0\right) \right] + \mathbb{P}_X\left[\|X\|^2 \right].$$

- Same bound as before for $\mathcal{L}(\mu) - \mathbb{P}_X\left[\|X\|^2\right]$.

Application to images: a small example

How to create patches with a random support ?

- Let $(X_i, i \in I) \in \mathbb{R}^I$ be a random image, where $|I| < \infty$ is the number of pixels.
- Represent the pixel location i by a random variable S, putting

$$\mathbb{P}_{S,V|X} = \frac{1}{|I|} \sum_{i \in I} \delta_i \otimes \delta_{X_i}, \quad \text{where } (S, V) \in I \times \mathbb{R}.$$

- Add noise, introducing $V' = V + \xi$ such that $\mathbb{E}\big(V' | X, S\big) = V$.
- Put $U = (S, V')$ and change the representation of X to $\mathbb{P}_{U|X}$.
- Use an auxiliary set of images represented by the distribution $Q_{\theta,U} \in \mathcal{M}_+^1\big[\Theta \times \big(I \times \mathbb{R}\big)\big]$, where $\big|\mathrm{supp}\big(Q_{\theta,U}\big)\big| < \infty$. Take for instance the empirical distribution of n independent copies of X, or more precisely $Q_{\theta,U} = \frac{1}{n} \sum_{j=1}^{n} \delta_j \otimes \mathbb{P}_{S,V|X=X_j}$, where $(X_j, 1 \leq j \leq n) \sim \mathbb{P}_X^{\otimes n}$.
- Solve $\inf_{\ell_\theta : \mathrm{supp}\left(Q_{U|\theta}\right) \to \{1,\dots,k\}} \inf_{Q_{X|\theta,\ell_\theta(U)}} Q_{\theta,U}\Big[\mathcal{K}\big(Q_{X|\theta,\ell_\theta(U)}, \mathbb{P}_{X|U}\big)\Big]$.

Application to images: a small example

- Define the patch process as $\mathbb{P}_{T|X} = Q_{\theta,\ell_\theta(U)|X}$.
- In the exact case where $\mathbb{P}_{X|U} = Q_{X|\theta,\ell_\theta(U)}$,

$$\frac{\mathrm{d}\mathbb{P}_{U|X,U\in\mathrm{supp}(Q_U)}}{\mathrm{d}\mathbb{P}_{U|U\in\mathrm{supp}(Q_U)}}(u) = Z_X^{-1} Q_{\theta|U=u}\left[\frac{\mathrm{d}Q_{\theta,\ell_\theta(U)|X}}{\mathrm{d}Q_{\theta,\ell(U)}}(\theta,\ell_\theta(u))\right].$$

- Cluster the patches solving

$$\inf_{\ell:\mathrm{supp}(R_T)\to\{1,\dots,k\}} \inf_{R_{X|\ell(T)}} R_T\left[\mathcal{K}\left(R_{X|\ell(T)}, \mathbb{P}_{X|T}\right)\right]$$

- Define a new representation as $\mathbb{P}_{W|X} = R_{\ell(T)|X}$.
- In the exact case where $R_{X|\ell(T)} = \mathbb{P}_{X|T}$ and $\mathrm{supp}(R_T) = \mathrm{supp}(\mathbb{P}_T)$,

$$\frac{\mathrm{d}\mathbb{P}_{T|X}}{\mathrm{d}\mathbb{P}_T}(t) = Z_X^{-1}\frac{\mathrm{d}R_{\ell(T)|X}}{\mathrm{d}R_{\ell(T)}}(\ell(t)),$$

showing that the previous representation $\mathbb{P}_{T|X}$ can be recovered exactly from the next one $\mathbb{P}_{W|X} = R_{\ell(T)|X}$ and the marginal distributions $\mathbb{P}_T$ and $R_{\ell(T)}$.

Application to images: a small example

Application to images: a small example

Figure : Extracted patches : 500 $\times$ 500 from two images 1000 $\times$ 1500. .

the training sample corresponds to the extracted patches.

Application on images: small example

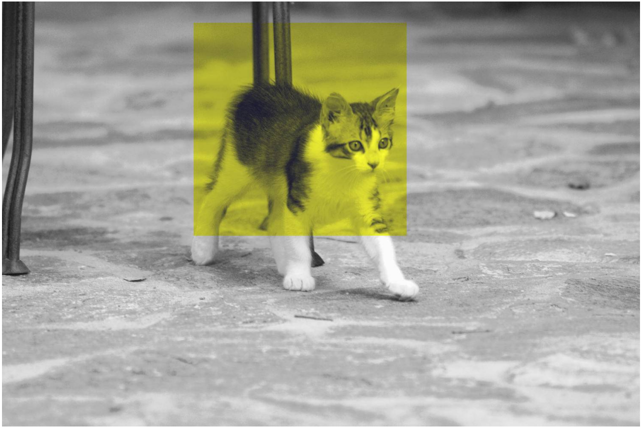

Figure : Selected image.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.

Application on images: small example

Figure : Clustering with information k-means.