Examples are not Enough, Learn to Criticize! Criticism for Interpretability

Been Kim, Rajiv Khanna, Oluwasanmi Koyejo

Wittawat Jitkrittum

Gatsby Machine Learning Journal Club
16 Jan 2017

Summary

Examples are not Enough, Learn to Criticize! Criticism for Interpretability
Been Kim, Rajiv Khanna, Oluwasanmi O. Koyejo
NIPS 2016.

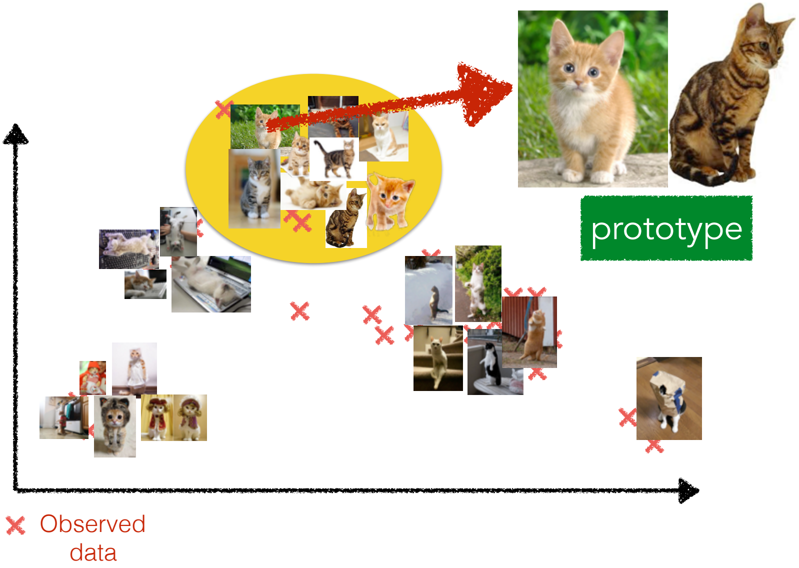
Given a big dataset, want to do 2 things:

1. **Summarize**: Find typical examples = prototypes. Majorities.
2. **Criticize**: Find atypical examples that are not covered by the prototypes. Minorities.

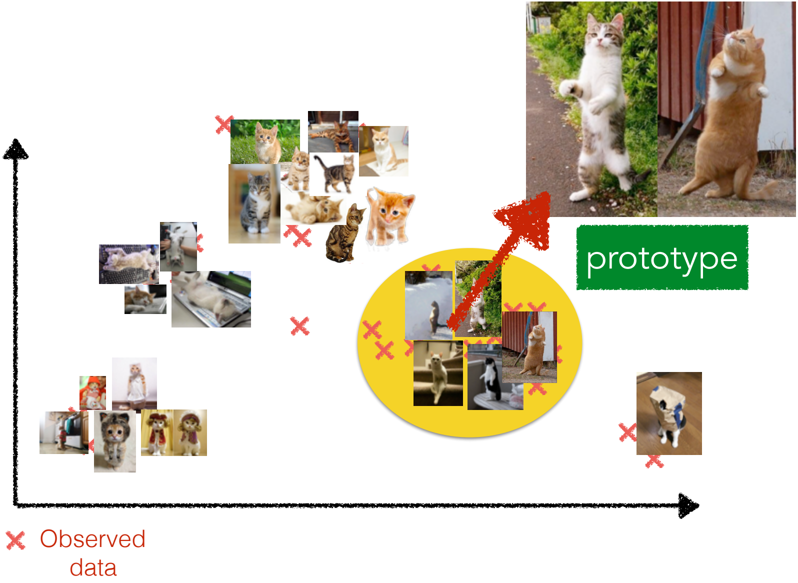
- Many existing works focus on only [1] e.g., K-medoid, set cover.
- Main message: [2] is also important.
- Use kernel MMD as the objective.

(Some slides are stolen from Been Kim.)

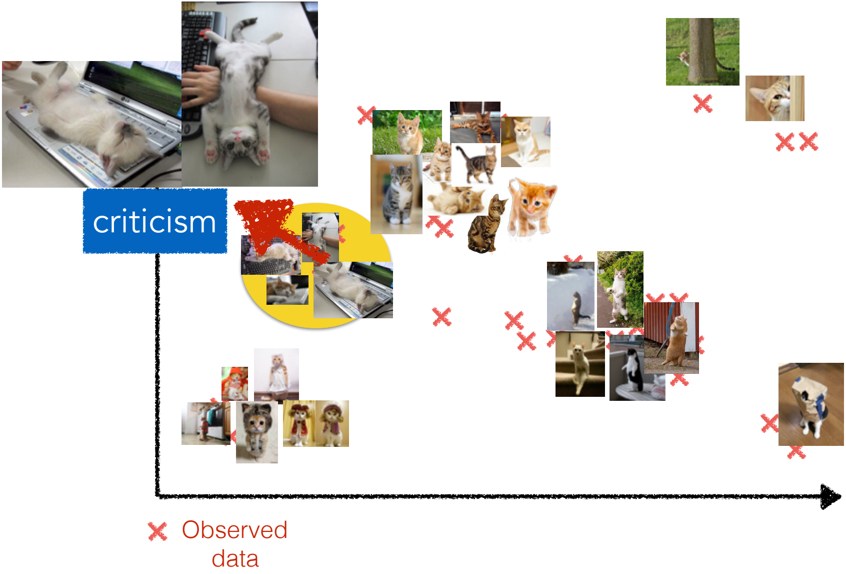
Understanding data through examples



Understanding data through examples



Understanding data through examples



Maximum Mean Discrepancy (MMD)

- k : a kernel associated with RKHS $\mathcal{H}$ s.t. $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.
- Two sets of samples: $X = \{x_i\}_{i=1}^{n} \sim P$, $Z = \{z_i\}_{i=1}^{m} \sim Q$.
- Empirical MMD:

$$\begin{aligned}
\mathrm{MMD}^2(X, Z) &= \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(z_j) \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, z_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(z_i, z_j).
\end{aligned}$$

- **Summarization**: Choose subset indices $\mathsf{S} \subset \{1, \ldots, n\}$ to minimize $\mathrm{MMD}^2(X, X_{\mathsf{S}})$. (Cf. kernel herding).
  - Pick $|\mathsf{S}| = m$ points to preserve the moments as defined by $\phi(\cdot)$.

Proposal: MMD-critic for Prototypes

- Given $X = \{x_i\}_{i=1}^n$, define a maximization objective $J_b(\mathsf{S})$

$$J_b(\mathsf{S}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) - \mathrm{MMD}^2(X, X_S)$$

$$= \underbrace{\frac{2}{n|\mathsf{S}|}\sum_{i=1}^{n}\sum_{j\in\mathsf{S}} k(x_i, x_j)}_{\text{relevancy}} - \underbrace{\frac{1}{|\mathsf{S}|^2}\sum_{i\in\mathsf{S}}\sum_{j\in\mathsf{S}} k(x_i, x_j)}_{\text{redundancy}}$$

- $\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)$ is constant. Added so that $J_b(\emptyset) = 0$ ("normalized").
- Select m_* prototypes by (discrete optimization)

$$\max_{\mathsf{S}\subset\{1,\ldots,n\}, |\mathsf{S}|\le m_*} J_b(\mathsf{S}).$$

Optimization Guarantees

- **Def**: $F(\mathsf{S})$ is normalized if $F(\emptyset) = 0$.
- **Def**: $F(\mathsf{S})$ is monotonic if $\mathsf{U} \subseteq \mathsf{V} \subseteq \{1, \ldots, n\}$ implies $F(\mathsf{U}) \leq F(\mathsf{V})$.
- **Def**: $F(\mathsf{S})$ is submodular if for all $\mathsf{U}, \mathsf{V} \subseteq \{1, \ldots, n\}$,

$$F(\mathsf{U} \cup \mathsf{V}) + F(\mathsf{U} \cap \mathsf{V}) \leq F(\mathsf{U}) + F(\mathsf{V}).$$

- Will show that $J_b(\mathsf{S})$ is monotonic, submodular under some conditions.
- Then, use greedy forward search. At each iteration t,

$$\mathsf{S}_{t+1} = \mathsf{S}_t \cup \{\arg \max_{u \in \{1,\ldots,n\} \setminus \mathsf{S}_t} J_b(\mathsf{S}_t \cup \{u\})\}.$$

Theorem (Nemhauser et al. (1978))

If F is normalized, monotonic, submodular, then the greedy approach achieves at least $(1 - e^{-1}) \max_{|\mathsf{S}| \leq m_*} F(\mathsf{S})$.

Variational View of MMD

$$\begin{aligned}\mathrm{MMD}(P, Q) &= \left\| \mathbb{E}_{X \sim P}[\phi(X)] - \mathbb{E}_{Y \sim Q}[\phi(Y)] \right\|_{\mathcal{H}} \\ &= \sup_{f \in \mathcal{H}, \|f\| \leq 1} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)].\end{aligned}$$

- $\arg\sup$ is the witness function:

$$f(x) = \mathbb{E}_{X' \sim P}[k(x, X')] - \mathbb{E}_{Y' \sim Q}[k(x, Y')].$$

- $f(x) > 0$ in high density areas of P.
- $f(x) < 0$ in high density areas of Q.
- Magnitude $|f(x)|$ indicates the density difference at x.

- For our purpose, the empirical witness associated with $\mathrm{MMD}(X, X_{\mathsf{S}})$:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} k(x, x_i) - \frac{1}{|\mathsf{S}|} \sum_{j \in \mathsf{S}} k(x, x_j).$$

MMD-critic for Criticisms

- **Criticisms of** S are points with high magnitude of the witness f

$$\mathsf{C} = \arg \max_{\mathsf{C} \subseteq \{1,\ldots,n\} \setminus \mathsf{S}, |\mathsf{C}| < c_*} L(\mathsf{C}) + \log \det K_{\mathsf{C},\mathsf{C}}$$

$$L(\mathsf{C}) = \sum_{l \in \mathsf{C}} |f(x_l)| = \sum_{l \in \mathsf{C}} \left| \frac{1}{n} \sum_{i=1}^{n} k(x_i, x_l) - \frac{1}{|\mathsf{S}|} \sum_{j \in \mathsf{S}} k(x_j, x_l) \right|.$$

- Regularizer $\log \det K_{\mathsf{C},\mathsf{C}}$ is high when $\{x_l\}_{l \in \mathsf{C}}$ are diverse.
- $L(\mathsf{C}) + \log \det K_{\mathsf{C},\mathsf{C}}$ is sub-modular. Greedy optimization.

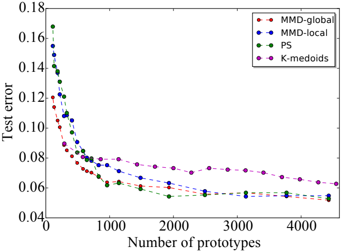
- The whole procedure gives summary points S, and criticisms C.

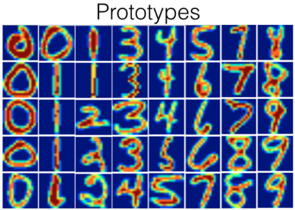
Quality of the Prototypes

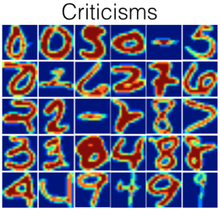
- Find prototypes of USPS handwritten digits.
- Gaussian kernel: $k(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$.
- Use 1-NN (nearest prototype) classification error as the quality measure.
- Let $y_i \in \{1, \ldots, 10\}$ be the class label of x_i.
- Given $\hat{x}$, the nearest prototype classifier predicts y_{i^*}, where

$$i^* = \arg\min_{i \in \mathsf{S}} \|\phi(\hat{x}) - \phi(x_i)\|_{\mathcal{H}}^2 = \arg\min_{i \in \mathsf{S}} k(\hat{x}, x_i).$$

Performance on USPS Data



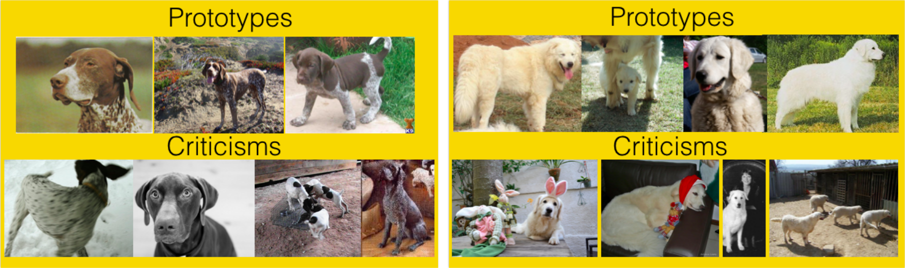




- MMD-local: Use $\exp(-\gamma||x_i - x_j||^2)[y_i = y_j]$. Supervised kernel to find the prototypes.
- MMD-global: Use the usual Gaussian kernel.
- PS: Prototype Selection of Bien and Tibshirani, 2011.
- Features = raw pixels.

Qualitative Measure: Prototype and Criticisms

- Two types of dog breeds from Imagenet.
- Features = image embeddings from He et al., 2015.



Pilot study with human subjects

- Definition of interpretability: A method is interpretable if a user can correctly and efficiently predict the method's results.

- Task: Assign a new data point to one of the groups using 1) all images 2) prototypes 3) prototypes and criticisms 4) small set of randomly selected images

a new data point

group 1

group 2

Eval3

Pilot study with human subjects

- Definition of interpretability: A method is interpretable if a user can correctly and efficiently predict the method's results.

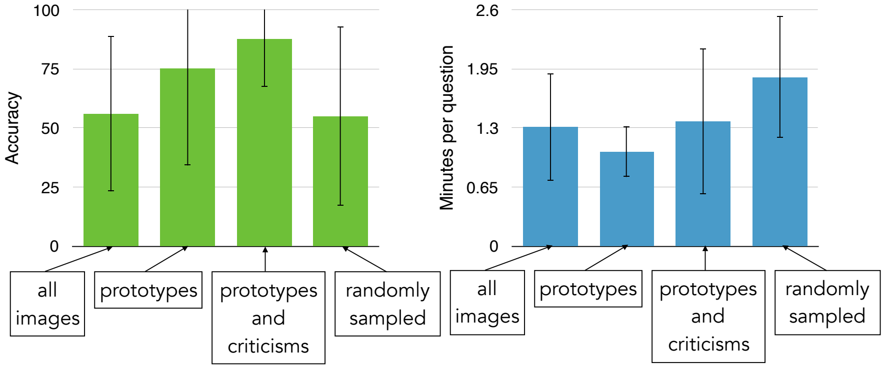
- Task: Assign a new data point to one of the groups using 1) all images 2) prototypes 3) prototypes and criticisms 4) small set of randomly selected images

a new data point

group 1

group 2

Pilot study with human subjects



Comment:

"[Proto and Criticism Condition resulted in] less confusion from trying to discover hidden patterns in a ton of images, more clues indicating what features are important"

n = 3
21 questions each

Some Questions

- What happens when there is no $\log\det K_{C,C}$?
- Quantify the effect of the image embeddings from He et al., 2015.
- There are only 3-4 human subjects.
- Possible to do a continuous optimization without selecting a subset?

Lemma 1: $J_b(\mathsf{S})$ Is Linear in K

- Let $K \in \mathbb{R}^{n\times n}$ such that $k_{ij} = k(x_i, x_j)$.
- Prototype objective:

$$
\begin{aligned}
J_b(\mathsf{S}) &= \frac{2}{n|\mathsf{S}|}\sum_{i=1}^{n}\sum_{j\in\mathsf{S}} k(x_i,x_j) - \frac{1}{|\mathsf{S}|^2}\sum_{i\in\mathsf{S}}\sum_{j\in\mathsf{S}} k(x_i,x_j) \\
&= \frac{2}{n|\mathsf{S}|}\sum_{i=1}^{n}\sum_{j=1}^{n}[j\in\mathsf{S}]k(x_i,x_j) - \frac{1}{|\mathsf{S}|^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[i\in\mathsf{S}][j\in\mathsf{S}]k(x_i,x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{2}{n|\mathsf{S}|}[j\in\mathsf{S}] - [i\in\mathsf{S}][j\in\mathsf{S}]\right)k_{ij} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\mathsf{S})k_{ij} := \langle A(\mathsf{S}), K\rangle\,,
\end{aligned}
$$

- Matrix inner product: $\langle A, B\rangle = \sum_i \sum_j a_{ij}\, b_{ij}\,.$

Theorem 2.1: Monotone Linear Forms

- Given $H \in \mathbb{R}^{n\times n}$ s.t. $0 \leq h_{ij} \leq h_*$ where $h_* := \max_{i,j} h_{ij} > 0$.
- Define $E \in \{0,1\}^{n\times n}$ s.t. $e_{ij} = [h_{ij} = h_*]$.
- Define $F(B, \mathsf{S}) := \langle A(\mathsf{S}), B\rangle$.
- Let $m := |\mathsf{S}|$. Define

$$\alpha(n,m) = \frac{F(E, \mathsf{S} \cup \{u\}) - F(E, \mathsf{S})}{F(1 - E, \mathsf{S})},$$

  for all $u \in \mathsf{S}$.
- If for all i, j s.t. $[h_{ij} \neq h_*]$, for all $m \in \{0, \ldots, n\}$, $h_{i,j} \leq h_* \alpha(n,m)$, then $F(H, \mathsf{S})$ is monotone.

- A similar statement to guarantee that $F(H, \mathsf{S})$ is submodular.

(Corollary) Monotone Submodularity for MMD

Assume

1. K is s.t., $k_{ij} \geq 0$.
2. $k_{i,i} = k_* > 0$ for all $i \in \{1, \ldots, n\}$.
3. K is diagonally dominant i.e., $\sum_{j \neq i} |k_{i,j}| < |k_{i,i}|$ for all i.
4. $k_{i,j} \leq \frac{k_*}{n^3+2n^2-2n-3}$

Then, $J_b(\mathsf{S})$ is monotone submodular.

- For a fixed n, and $k_{i,j} = \exp\left(-\gamma \|x_i - x_j\|^2\right)$, there exists γ such that (3), (4) are satisfied.
- What if n is very large?

Questions?

Thank you