Nonparametric Kernel Estimators for Image Classification

Barnabás Póczos, **Liang Xiong**, **Dougal J. Sutherland**, and **Jeff Schneider**
Auton Lab, School of Computer Science, Carnegie Mellon University – autonlab.org

Motivation

The "bag of features" approach to image classification, representing images as a set of local features over a grid, is very powerful and popular.

But to use it for learning methods like SVMs, we need either a mapping into R^n (e.g. "bag of words" or histograms) or a kernel directly on these sets.

We construct such a kernel by considering each sets as a samples from an unknown probability distribution, then nonparametrically estimating divergences between the distributions.

Support Distribution Machines

Many kernel functions can be computed from:

$$D_{\alpha,\beta}(p\|q) = \int p(x)^\alpha \, q(x)^\beta \, p(x) \, \mathrm{d}x$$

including linear $\int pq$, polynomial $(c + \int pq)^s$, and Gaussian $\exp\left(-\frac{1}{2}\mu^2(p,q)/\sigma^2\right)$, where we can use various "distances" μ:

$$L_2 : \mu^2 = \int p^2 + \int q^2 - 2\int pq$$
$$\text{Rényi-}\alpha : \mu_\alpha = \log\left(\int p^\alpha q^{1-\alpha}\right)/(\alpha - 1)$$
$$\text{Tsallis-}\alpha : \mu_\alpha = \left(\int p^\alpha q^{1-\alpha} - 1\right)/(\alpha - 1)$$
$$\text{Hellinger} : \mu^2 = 1 - \int \sqrt{pq}$$
$$\text{Bhattacharyya} : \mu^2 = -\log \int \sqrt{pq}$$

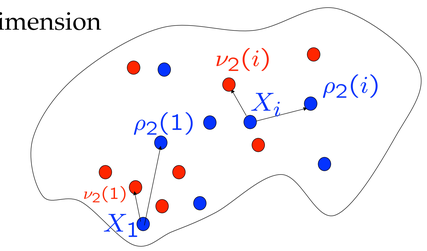
KL divergence is the limit of Rényi-α as $\alpha \to 1$.

To get a kernel matrix, we:
1. Estimate D matrix (pairwise from samples)
2. Plug into the formulae above to get K
3. Symmetrize: $K := (K + K^T)/2$
4. Project to PSD: discard negative eigenvalues

Divergence Estimation

We estimate D with k^{th}-nearest-neighbor distances:

- $X_{1:n}$: n samples from p; $Y_{1:m}$: m samples from q
- $\rho_k(i)$: the distance to the k^{th} neighbor of X_i in $X_{1:n}$
- $\nu_k(i)$: the distance to the k^{th} neighbor of X_i in $Y_{1:m}$
- d: dimension

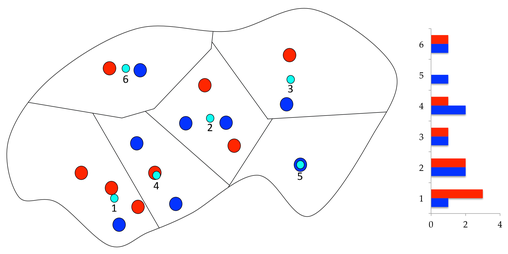


$$\widehat{D}_{\alpha,\beta}(X_{1:n}, Y_{1:m}) = \frac{B_{k,d,\alpha,\beta}}{n\,(n-1)^\alpha\, m^\beta} \sum_{i=1}^{n} \rho_k^{-d\alpha}(i)\,\nu_k^{-d\beta}(i)$$

$$B_{k,d,\alpha,\beta} = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \, \frac{\Gamma(k)^2}{\Gamma(k-\alpha)\,\Gamma(k-\beta)}$$

For *fixed* k (we use 5), this estimator is provably L_2 consistent and asymptotically unbiased.

Comparison to Bag of Words



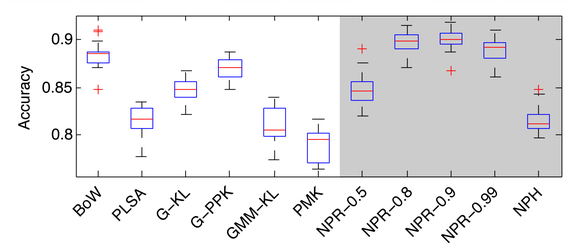
BoW loses information in quantization, including correlations between codewords, and requires tuning the codebook size.

Image Classification

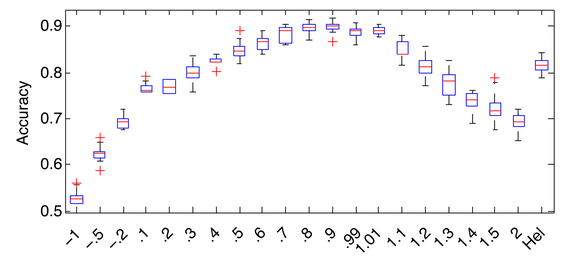
We will now show experimental results on classifying images into categories, based on 384-dimensional color SIFT features after PCA dimensionality reduction. We compare to:

- Bag of words (BoW)
- BoW processed by pLSA [1]
- the Pyramid Matching Kernel (PMK) [2]
- Kernels based on Gaussians and GMMs:
  - with KL divergence [3]
  - with Probability Product Kernels [4]

Object Classification (ETH-80)



16 runs, 10-fold CV; 18 PCA dimensions.

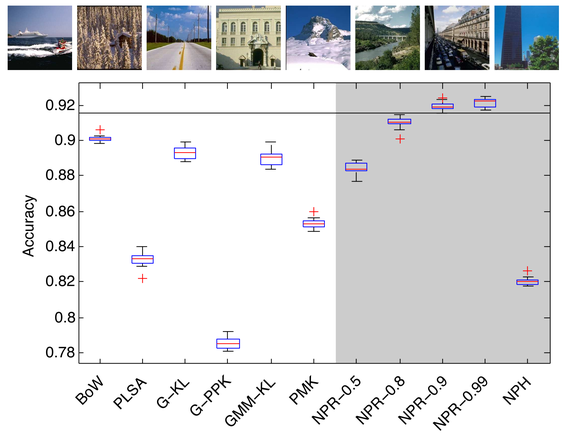


Performance by α value on ETH-80: values near but slightly below 1 seem best.

Acknowledgements

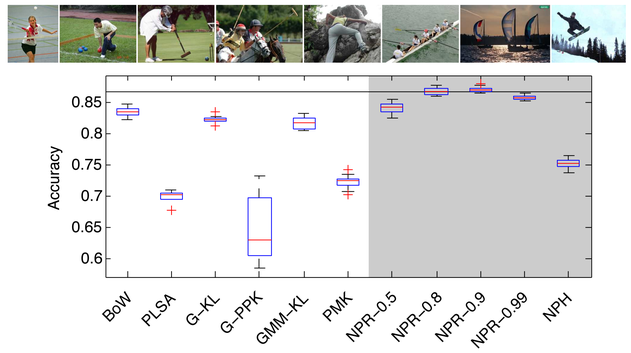
This work was funded in part by the National Science Foundation under grant NSF-IIS0911032 and the Department of Energy under grant DESC0002607.

Scene Classification (Oliva/Torralba)



16 runs, 10-fold CV; 53 PCA dimensions, plus y coordinate. Beats best previous result [5].

Sport Classification (Li/Fei-Fei)



16 runs, 2-fold CV; 57 PCA dimensions, plus x, y coordinates. Matches best previous result [6].

Takeaway

The bag of features model is powerful. But in quantizing it, we lose some of that power.

We can improve performance by using the same features with better dissimilarity measures.

The nonparametric divergence estimator presented here matches or beats state-of-the-art techniques using learned features.

References

[1] T. Hofmann. Probabilistic latent semantic analysis. UAI, 1999.
[2] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. JMLR, 2007.
[3] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. NIPS, 2004.
[4] T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. JMLR, 2004.
[5] J. Qin and N. H. Yung. SIFT and color feature fusion using localized maximum-margin learning for scene classification. ICMV, 2010.
[6] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition. CVPR, 2011.