Learning Efficient Random Maximum A-Posteriori Predictors with Non-Decomposable Loss Functions

Tamir Hazan (University of Haifa) Subhransu Maji (TTI Chicago) Joseph Keshet (Bar-Ilan University) Tommi Jaakkola (MIT)

Motivation

We investigate the Bayeisan aspects of PAC-Bayesian generalization bounds.

Contributions: Posterior distributions that allow efficient sampling procedures

- Posterior distributions for supermodular predictions.
- Predictive models for approximate inference / LP relaxations.
- Empirical risk minimization for any loss function and smooth posterior.

Background

Supervised learning: given training data (x,y) ∈ S, learn parameters w to derive prediction rule $y_w(x)$ that minimizes the risk.

- Maximum A-Posteriori (MAP) prediction:

$$y_w(x) = \arg\max_{y_1,\ldots,y_n} \theta(y;x,w)$$

- Random MAP predictor:

$$p[y|x] = P_{\gamma\sim q_w}\left[y = y_\gamma(x)\right]$$

- Bayesian risk:

$$R(w,x,y) = \sum_{\hat{y}} P[y|x]L(\hat{y},y)$$

$$R(w) = E_{(x,y)\sim\rho}\left[R(w,x,y)\right]$$

$$R_S(w) = \frac{1}{|S|}\sum_{(x,y)\in S} R(w,x,y)$$

PAC-Bayesian generalization

For any δ and any λ>0, with probability at least 1-δ over the draw of the training set, the following holds simultaneously for all w:

$$R(w) \le \frac{1}{1-\exp(-\lambda)}\left(\lambda R_S(w) + \frac{KL(q_w||p)+\log(1/\delta)}{|S|}\right)$$

while $KL(q_w||p) = \int q_w(\gamma)\log(q_w(\gamma)/p(\gamma))$

When we can minimize risk?

To find the best parametrized posterior distribution $q_w(\gamma)$ we minimize the bound, as long as the posterior is smooth

$$\nabla_w R(w,x,y) = E_{\gamma\sim q_w}\left[\nabla_w[\log q_w(\gamma)]L(y_\gamma(x),y)\right]$$

$$\nabla_w KL(q_w||p) = E_{\gamma\sim q_w}\left[\nabla_w[\log q_w(\gamma)]\left(\log(q_w(\gamma)/p(\gamma))+1\right)\right]$$

Proof: $R(w,x,y) = \int q_w(\gamma)L(y_\gamma(x),y)d\gamma$

Differentiate under the integral and use

$$\nabla_w q_w(\gamma) = q_w(\gamma)\nabla_w \log(q_w(\gamma))$$

Priors set regularizations

The Complexity of the bound (regularization) is determined by its prior distribution:

Let $q_w(\gamma) = q_0(\gamma - w)$ then

$$KL(q_w||p) = -H(q_0) - E_{\gamma\sim q_0}[\log p(\gamma+w)]$$

For Gaussian prior $\nabla_w KL(q_w||p) = w$

Proof: Change variable $\hat{\gamma} = \gamma - w$

Learning efficient posteriors

Learn supermodular MAP predictors

$$\theta(y;x,w) = \sum_{i\in V}\theta_i(y_i;x,w) + \sum_{i,j\in E}\theta_{i,j}(y_i,y_j;x,w)$$

$$\theta_{i,j}(y_i,y_j;x,w) = w_{i,j}y_iy_j, \quad w_{i,j}\ge 0$$

Multiplicative posteriors result in log-barrier functions over the parameters: For any probability distribution $q_1(\gamma)$ over the nonnegative reals, let $q_w(\gamma) = q_1(\gamma/w)/w$

$$KL(q_{\alpha,w}||p) = -H(q_\alpha) - \log w - E_{\gamma\sim q_\alpha}[\log p(w\gamma)]$$

Proof: Change of variable

$$-H(q_w) \overset{\hat{\gamma}=\gamma/w}{=} \int q_1(\hat{\gamma})\left(\log q_1(\hat{\gamma}) - \log w\right)d\hat{\gamma} = -H(q_1) - \log w.$$

For Gaussian prior: $E_{\gamma\sim q_1}[\log p(w\gamma)] = -\frac{1}{2}w^2 + c$

For exponential prior: $E_{\gamma\sim q_1}[\log p(w\gamma)] = -w$

Learn with approximate MAP prediction

$$b^* \in \arg\max_{b_r(y_r)} \sum_{r,y_r} b_r(y_r)\theta_r(y_r;x,w)$$

$$s.t.\ \ b_r(y_r)\ge 0,\ \sum_{y_r} b_r(y_r) = 1,\quad \sum_{y_s\setminus y_r} b_s(y_s) = b_r(y_r)\quad \forall r\subset s$$

Any optimal solution b* is described by

$$\tilde{y}_w(x) = (\tilde{y}_{w,r}(x))_{r\in\mathcal{R}} \ \text{ where }\ \tilde{y}_{w,r}(x) = \{y_r : b_r^*(y_r) > 0\}$$

Proof: Any feasible solution that has the same support as b* is also optimal solution. Follows from Lagrange optimality conditions

$$\sum_r \max_{y_r}\left\{\theta_r(y_r;x,w) + \sum_{c:c\subset r}\lambda_{c\to r}(y_c) - \sum_{p:p\supset r}\lambda_{r\to p}(y_r)\right\}$$

Empirical Evaluation

Learning supermodular segmentations with non-decomposable loss functions (Grabcut)

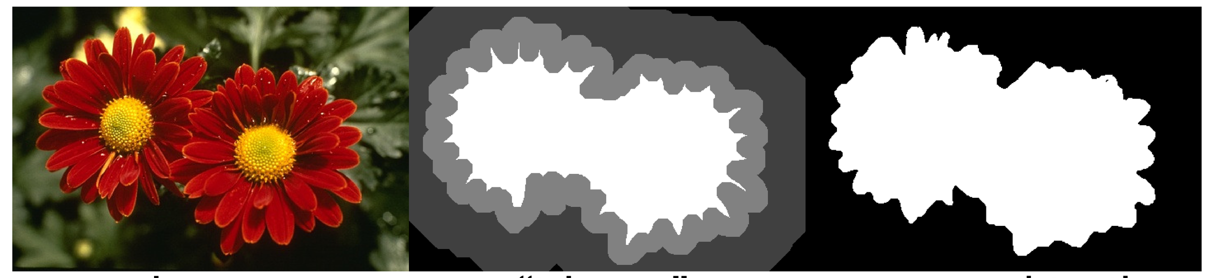


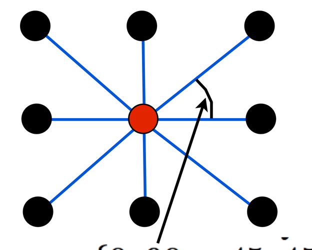
image "trimap" ground truth

$$\theta(y;x,w) = \sum_{i\in V}\theta_i(y_i;x,w) + \sum_{i,j\in E}\theta_{i,j}(y_i,y_j;x,w)$$

$$\theta_i(y_i;x,w) = w_{y_i}\log P(y_i|x),\quad w_{y_i}\in R$$

$$\theta_{i,j}(y_i,y_j;x,w) = w_a\exp(-(x_i-x_j)^2)y_iy_j,\quad w_a \ge 0$$

Unary potentials are obtained using color Gaussian mixture models learned from the initial "trimap".



angles $a \in \{0, 90, -45, 45\}$

Measuring segmentation loss

Given a segmentation

$$A[i] \in \{0 \text{ background}, 1 \text{ foreground}\}$$

$$\text{GrabcutLoss(A,B)} = \frac{\sum_{i\in U}\mathbb{1}(A[i]\ne B[i])}{|U|}$$ ← unknown region

measures incorrect pixels

$$\text{PASCALLoss(A,B)} = 1 - \frac{\sum_i \mathbb{1}(A[i]\otimes B[i])}{\sum_i \mathbb{1}(A[i]\oplus B[i])}$$

measures pixel overlap (set intersection over union)

Method	Grabcut loss	PASCAL loss
Our method	**7.77%**	**5.29%**
Structured SVM (hamming loss)	9.74%	6.66%
Structured SVM (all-zero loss)	7.87%	5.63%
GMMRF (Blake et al. [1])	7.88%	5.85%
Perturb-and-MAP ([17])	8.19%	5.76%

Results on the Grabcut dataset (Blake et. al., ECCV 04)