Logistic Regression

per-pixel average negative log marginal error rate

<table>
<thead>
<tr>
<th>Error rate</th>
<th>24.5</th>
<th>24.6</th>
<th>24.7</th>
<th>24.8</th>
<th>24.9</th>
<th>25.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>CRF LBMO + bagging</td>
<td>CRF LBMO bag 50 WL</td>
<td>CRF LBMO bag 15 WL</td>
<td>GPstruct 15 WL</td>
<td>GPstruct 50 WL</td>
<td></td>
</tr>
</tbody>
</table>

Bayesian Learning & prediction (standard GPstruct)

Output structure imposed by Markov random field:

$p(y|x, f) = \frac{\exp \sum_{i \in \text{pixel positions}} \kappa(y_i, y'_i)}{Z(y|x, f)}$

The factors of the MRF are jointly given a Gaussian prior process:

$f \sim \mathcal{GP}(0, \kappa(\cdot, \cdot))$

Kernel design in this paper:

$\kappa(c_{\text{unary}}, c_{\text{unary}}) = 0$
$\kappa(c_{\text{binary}}, (y_i, y_{i+1}), c_{\text{unary}}) = (g_k(x_i, x_{i+1}))$

Training: intractable likelihood

likely needed for MCMC: grid MRF intractable due to global normalization. Solution: pseudo-likelihood approximation, locally normalized: $p(y|f) = \prod_{p=1}^P p(y_p|y_{p+1}, \ldots, y_P, f) \approx \prod_{p=1}^P p(y_p|\text{neighbours}(p), f)$

Prediction: intractable marginals

Pixel-wise MAP intractable in grid MRF. Solution: Obtain approximate marginals using tree-reweighted belief propagation.

Training: kernel matrix size

How many factors $f(x, y_p)^2$ One for each $p$ in $\Delta_{train}$, $y_p \in \mathcal{Y}$, ie $(10^5)^2 \times 10^5$. So $K$ square of shape $10^5 \times 10^5$. Solution: ensemble learning, split up training pixels over weak learners.

Experimental setup

Two image segmentation datasets, all resized to 50*150 pixels, 8 or 9 classes.

Compare:

- GPstruct+bagging
- CRF trained using PL
- CRF trained using loss-based marginal optimisation (Domke 2013)
- CRF LBMO+bagging

Learning & prediction (standard GPstruct)

$p(f|D) \propto p(f) \prod_{(x,y) \in D} p(y|x, f)$

Sampling from $f|D$ not analytical (non-Gaussian likelihood): elliptical slice sampling.

$p(y^*|x^*, D) = \int \int p(y^*|x^*, f^*) p(f^*|f)p(f|D)df^*df$

Using pixel-wise Hamming error, hence need posterior marginals $p(y_p|x^*, D)$.

Challenges

Training: intractable likelihood

likely needed for MCMC: grid MRF intractable due to global normalization. Solution: pseudo-likelihood approximation, locally normalized: $p(y|f) = \prod_{p=1}^P p(y_p|y_{p+1}, \ldots, y_P, f) \approx \prod_{p=1}^P p(y_p|\text{neighbours}(p), f)$

Prediction: intractable marginals

Pixel-wise MAP intractable in grid MRF. Solution: Obtain approximate marginals using tree-reweighted belief propagation.

Training: kernel matrix size

How many factors $f(x, y_p)^2$ One for each $p$ in $\Delta_{train}$, $y_p \in \mathcal{Y}$, ie $(10^5)^2 \times 10^5$. So $K$ square of shape $10^5 \times 10^5$. Solution: ensemble learning, split up training pixels over weak learners.

Experimental results

Question: does our method give good performance? Bagging also helps CRF LBMO, but GPstruct still remains better. Hence bagging alone does not explain GPstruct’s performance.

Question: effect of PL+TRW approximation? New experiment: compare exact and approx versions of likelihood/prediction function. New dataset allowing exact computations:

- scaled down images of Stanford dataset (10*10 pixels)
- reduced to foreground-background segmentation
- training set size 100, test set size 100, 4 shuffles

TRW produces excellent approximations. When used inside ESS in GPstruct, PL produces a robust approximation.

Future work

- variational Bayes; requires bounding intractable sums
- inducing points, GP sparsification
- kernel hyperparameter learning
- matrix factorization on $K$
- richer/structured kernels