Relevance Vector Machines & Eponine Algorithm

Konstantin Tretjakov (kt@ut.ee)

22. november 2005

Contents

- RVM machine learning algorithm
- Eponine models: EAS, EWS, C-EWS.
- Applications to genomic sequence analysis.

What & Why?

- The problem as usual: „we've got this DNA, but we don't know what to do with it".
- „We also have this nice machine learning method that looks cool. . . "
- The usual solution:
  - Make it applicable to sequence analysis
  - Apply it to sequence analysis
  - (Make up some cool-looking results)
  - Publish the results.
- And we here shall just observe, discuss and learn.

RVM

- Proposed by Tipping as a competitor to SVM.
- A kind of generalized linear model for regression:

$$y(\mathbf{x}) = \sum_m w_m \phi_m(\mathbf{x}) + w_0 + \epsilon$$

…or classification:

$$y(\mathbf{x}) = \sigma \left(\sum_m w_m \phi_m(\mathbf{x}) + w_0 \right)$$

RVM

- Assume a model with gaussian noise

$$P(\mathbf{t}|\mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{\Phi w}\|^2\right)$$

- We can now use the Bayes' rule:

$$P(\mathbf{w}|\mathbf{t}) = \frac{P(\mathbf{t}|\mathbf{w})P(\mathbf{w})}{P(\mathbf{t})}$$

- To avoid overfitting define *priors* for $P(\mathbf{w})$

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_i \mathcal{N}(w_i|0, \alpha_i^{-1})$$

RVM

- The parameters α_i specify the prior distribution of w_i. Now we must also estimate them somehow. Again, use some Bayes', some integration and some ugly calculations to get something like:

$$P(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = \frac{1}{(2\pi)^{(N+1)/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w} -$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\Phi} + \mathbf{A})^{-1}$, $\mu = \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{B} \mathbf{t}$,
$A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, $\mathbf{B} = \sigma^{-2} \mathbf{I}_n$.

- Now integrate away the weights to get a similarly ugly expression for $P(\mathbf{t}|\boldsymbol{\alpha})$.

RVM

- We can estimate $\boldsymbol{\alpha}$ by maximizing $P(\mathbf{t}|\boldsymbol{\alpha})$ by using, say, gradient descent.
- We could also just used the EM algorithm to maximize $P(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$ in the first place, using $\boldsymbol{\alpha}$ as hidden parameters.
- In both cases we get some algorithm that updates $\mathbf{w}$ and $\boldsymbol{\alpha}$ iteratively.
- During the updates many α-s will tend to ∞, so we can just prune the corresponding weights.
- As more weights get pruned, the algorithm runs progressively faster.

RVM for Classification

- For classification we apply the logistic function $\sigma(x) = \frac{1}{1+e^x}$ on the result of linear RVM. This lets us to interpret the value of the model as a probability:

$$y(\mathbf{x}) = \sigma\left(\sum_m w_m \phi_m(\mathbf{x}) + w_0\right)$$

- The probability of the data is then:

$$P(\mathbf{t}|\mathbf{w}) = \prod_i y(\mathbf{x})^{t_i}(1 - y(\mathbf{x}))^{1-t_i}$$

- This expression is worse than the previous and we can't integrate the weights away in closed form.

RVM for Classification

- Iterative algorithm by MacKay:
  - Fix some $\boldsymbol{\alpha}$. Find the maximum likelihood estimate for weights $\mathbf{w}_{ML}$.
  - Assume that the distribution $P(\mathbf{w}, \mathbf{t}|\boldsymbol{\alpha})$ of $\mathbf{w}$ is a gaussian centered at $\mathbf{w}_{ML}$.
  - Estimate the covariance matrix of this gaussian: $\nabla^2 \log P(\mathbf{w}, \mathbf{t}|\boldsymbol{\alpha}) =$ some decently simple expression.
  - Update $\boldsymbol{\alpha}$ using this approximation.
- They say it works.

RVM for Classification

- In the end we get the values w_i for the model, and most of the values will be zero, which means the corresponding basis functions ϕ_i are not relevant for classification/regression.
- The next question — how to apply it to sequence data?
- All we need to do is define some nice basis functions ϕ_i.

Eponine Anchored Sequence

- The EAS model is used to classify *points in sequences*.
- Example application: transcription start site classification.
- Input: A sequence chunk and a coordinate of a point in it.
- Output: a „score” of the given position.

Eponine Anchored Sequence

- EAS consists of *positioned constraints (PC)*
- PC = a distribution over integer offsets P + a PWM W.
- Score of a given input example C with anchor point a given by a PC ϕ is:

$$\phi(C) = \frac{1}{|W|} \log \sum_i P(a+i) W(C, a+i)$$

- EAS = a weighted sum of several PC-s.

Eponine Anchored Sequence

- Clearly, EAS is an instance of RVM with basis functions ϕ_i being the PC-s.
- So we can use RVM training to infer the parameters of the EAS model.
- Problem: the set of all PC-s is too large.
- Solution: Use *Sampling-RVM*

Sampling-RVM

- Start with some randomly chosen subset of basis functions.
- During training some of them will get pruned away.
- Once the working set gets too small, fill it up with new functions.
- New functions can be generated either randomly, or (better) using something similar to a genetic algorithm.

EAS applications

- Applied to TSS prediction.
- Input: EPD data.
- Output: Trained model.
- The model consisted of 4 basis functions.
    - One of them indicated the TATA box centered at -30.
    - The three others favored CG-rich regions
- The predictive performance of the model was comparable to existing methods (PromoterInspector, CpG): sensitivity 53%, specificity 73% (on a specially prepared pseudo-chromosome).

EWS

- Eponine Windowed Sequence: similar to EAS, but this time analyses subsequences, not just points.
- Consists of a PWM, that is scanned over the whole region.

$$\phi(S) = \frac{4^{|W|}}{|S| - |W| + 1} \sum_i W(S_{i \ldots i+|W|})$$

- Analogously, use the Sampling-RVM technique for model training.
- Generalizable to C-EWS.

Convolved-EWS

- Not a single PWM, but a scaffold of them: $W_1, W_2, \ldots, W_m$, each with associated position distribution $P_k(i)$.

$$\phi(S) = Z \sum_{i=n}^{m} \left(\prod_{k} \left(\sum_{j} P_k(j) W_k(S_{j\ldots j+|W_k|}) \right) \right)$$

  where $Z = 4^{\sum_k |W_k|}/m - n + 1$ and $n \ldots m$ is the length of the „window” corresponding to the scaffold.

- If scaffold contains only a single PWM, then it’s just EWS.

EWS/C-EWS applications

- Applied to
  - Predict promoter sites. Positive results, but EAS is better because it can point out the precise locations.
  - Analyze exonic splice sites.

Exonic splice site prediction

- Problem: detect splice sites within protein-coding sequence.
- Not easy, exactly because that's within protein-coding sequence.
- Solution: randomize the DNA sequence while preserving its amino-acid content and statistical properties. Then compare the randomized sequences with the originals for motifs.

Sequence randomization

- For each codon C attempt to substitute it with a synonym C'.
- If $P(C') > P(C)$ substitute the codon, otherwise substitute with probability $\frac{P(C')}{P(C)}$ (Metropolis-Hastings algorithm).
- In $P(C)$ take the context into consideration.
- Produces a randomized sequence with same statistical properties that codes the same amino-acid-chain.

Exonic splice sites: Results

- The trained model can distinguish randomized and non-randomized sequences sufficiently well.
- The model consists of 216 scaffolds!
- It's not clear why is it like that. Need more research.

Summary

- RVM is a (probably good) machine learning method. Might be even better than SVM.
- Eponine model makes it applicable to sequence analysis.
- If you are lucky you'll get useful information from the genome by applying this combination.

Questions?