Relevance Tagging Machine

Dmitry Molchanov
Dmitry Kondrashkin

10.04.2015

Model

We address the binary classification problem with binary features.

- Let $(X_i, T_i)_{i=1}^n$ be the training set
- $X_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ is an object and $T_i \in \{0, 1\}$ is class label
- $x_{ij} = 1 \Leftrightarrow X_i$ has the tag j
- All tags affect the class label independently

Probabilistic model of RTM:

$$q_j = \mathrm{P}(t = 1 | x_j = 1), \quad \mathrm{P}(t = 1 | x, q) = \frac{\prod_{j=1}^{d} q_j^{x_j}}{\prod_{j=1}^{d} q_j^{x_j} + \prod_{j=1}^{d} (1 - q_j)^{x_j}}$$

Feature selection

Bayesian *automatic relevance determination* (ARD) approach.

- Independent priors are placed over parameters q
- Hyperparameters are trained by maximizing the evidence

Symmetrical Beta distribution:

$$q_j \sim \mathrm{Beta}(q_j | \alpha_j + 1, \alpha_j + 1), \alpha_j \in [0, +\infty)$$

- $\alpha_j = 0 \Rightarrow q_j^{MAP} = q_j^{ML}$
- $\alpha_j = +\infty \Rightarrow q_j = 0.5 \Rightarrow q_j$ is removed from the model

Evidence

Bayes' theorem:

$$p(q|X, T, \alpha) = \frac{\mathrm{P}(T|X, q)p(q|\alpha)}{\int \mathrm{P}(T|X, q)p(q|\alpha)dq}$$

- $p(q|X, T, \alpha)$ – posterior
- $\mathrm{P}(T|X, q)$ – likelihood
- $p(q|\alpha)$ – prior
- $E = \int \mathrm{P}(T|X, q)p(q|\alpha)dq$ – evidence

Evidence can be used as a measure of model complexity.

Evidence maximization

Evidence is intractable as likelihood and prior are non-conjugate. Therefore we optimize it's approximation.

$$\tilde{E}(\alpha) \approx E(\alpha) \Rightarrow \arg\max_{\alpha} \tilde{E}(\alpha) \approx \arg\max_{\alpha} E(\alpha)$$

We consider two ways of approximation:

- Expectation Propagation
- Variational lower bounds

Expectation Propagation

Likelihood is approximated in the following form:

$$\mathrm{P}(T|X, q) \approx \frac{1}{Z} \prod_{j=1}^{d} q_j^{a_j} (1 - q_j)^{b_j}$$

Evidence approximation:

$$E(\alpha) \approx \tilde{E}(\alpha) = \frac{1}{Z} \int \prod_{j=1}^{d} \frac{q_j^{a_j+\alpha_j} (1 - q_j)^{b_j+\alpha_j}}{\mathrm{B}(\alpha_j + 1, \alpha_j + 1)} dq$$

$$\log \tilde{E}(\alpha) = -\log Z + \sum_{j=1}^{d} \log \frac{\mathrm{B}(a_j + \alpha_j + 1, b_j + \alpha_j + 1)}{\mathrm{B}(\alpha_j + 1, \alpha_j + 1)}$$

Hyperparameters optimization:

$$\alpha_j^* = \arg\max_{\alpha_j} (\log \mathrm{B}(a_j + \alpha_j + 1, b_j + \alpha_j + 1) - \log \mathrm{B}(\alpha_j + 1, \alpha_j + 1))$$

Variational lower bounds

$g(x, \eta)$ is called a variational lower bound of $f(x)$, if

$$f(x) \geq g(x, \eta) \; \forall x, \eta$$

$$f(x) = g(x, x) \; \forall x$$

We derive a variational lower bound for the likelihood of an object X_i:

$$\mathrm{P}(T_i | X_i, q) \geq L_i(q, \eta_i) = \prod_{j=1}^{d} L_{ij}(q_j, \eta_i), \;\; \eta_i \in [0, 1]^d$$

It gives us a family of evidence lower bounds:

$$E(\alpha) \geq \tilde{E}(\alpha, \eta) = \prod_{j=1}^{d} \int_0^1 \prod_{i=1}^{n} L_{ij}(q_j, \eta_i) p(q_j | \alpha_j) dq_j$$

EM algorithm

EM algorithm for evidence lower bound optimization:

1. E-step: $\eta^{new} = \arg\max_{\eta} \log \tilde{E}(\alpha^{old}, \eta)$
2. M-step: $\alpha^{new} = \arg\max_{\alpha} \log \tilde{E}(\alpha, \eta^{new})$

E-step still takes too much time, so we propose a simplification:

$$\eta_i^{new} = q^{MAP} = \arg\max_{q} \mathrm{P}(T|X, q)p(q|\alpha^{old}) \ \forall i = 1..n$$

Note that here $\eta_i = \eta_k \ \forall i, k = 1..n$.

Synthetic data experiments

500 objects, 50 tags

Percentage of removed noise features

Noise	MAP-EM	full EM	EP	RVM
random	90.87%	78.97%	89.1%	91.67%
correlated	75.88%	88.78%	51.15%	72.33%

Percentage of removed relevant features

Noise	MAP-EM	full EM	EP	RVM
random	1.69%	3.33%	0.67%	1.1%
correlated	1.83%	0.83%	0.33%	1.5%

Sentiment analysis

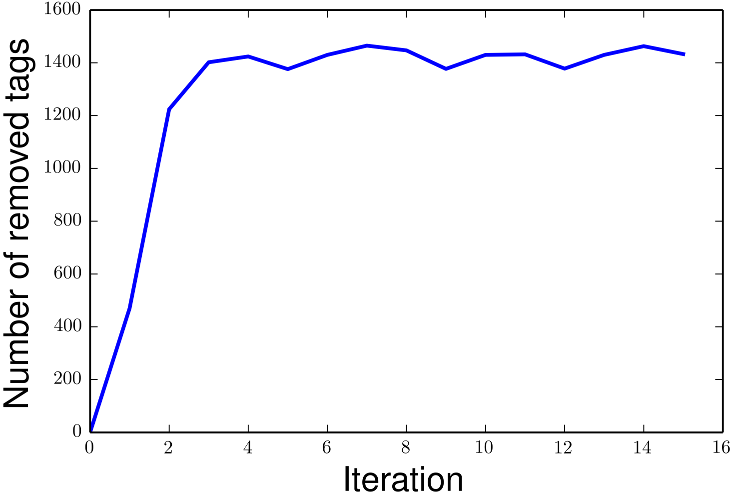
- 1000 train objects, 411 test objects, 1869 features
- Objects examples:
  "Brokeback Mountain is awesome."
  "Which answers why I dislike brokeback mountain..."
- 11 tags per object in average.
- Objects: stemming + bag of words

Classification accuracy:

MAP-EM	EP	RVM	LR	RF	GBDT	SVM
0.9659	0.9683	0.9586	0.9708	0.9416	0.9683	0.9683

LR – logistic regression
RF – random forest
GBDT – gradient boosting over decision trees (stumps)

EM tag removal



Conclusion

EP:

- Fast
- May fail to remove correlated tags
- Still proved to work well on real data

EM:

- Slow
- Provides better relevance determination
- Most irrelevant tags are removed on early steps

Both:

- Comparable to state of the art methods prediction accuracy
- Good feature selection on real data