

Lq-regularised sparse classifiers: A PAC-Bayes analysis

Ata Kaban (A.Kaban@cs.bham.ac.uk)

School of Computer Science, The University of Birmingham, B15 2TT, UK

Consider the sparse logistic regression over a training set of input-target pairs $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^n$, where $\mathbf{x}_j \in \mathbb{R}^m$ are m -dimensional data and $y_j \in \{-1, 1\}$ their labels:

$$\max_{\mathbf{w}} \sum_{i=1}^n -\log(1 + \exp(-y\mathbf{w}^T \mathbf{x}_i)) - \lambda \|\mathbf{w}\|_q^q \quad (1)$$

where $\|\mathbf{w}\|_q = (\sum_{i=1}^m |w_i|^q)^{1/q}$, with $q \leq 1$. When $q = 1$, we have L1-regularized logistic regression, studied e.g. in [5] and [1]. In turn, if $q \in (0, 1)$, we have a non-convex regularization term, which we will refer to as L_q -regularization. Parameter estimation becomes more difficult — though iterative algorithms that find local optima seem to be quite effective [2], and the concavity of the regularisation term may have beneficial effect in high dimensional problems.

For the analysis that follows, we observe that the optimisation problem (1) may be interpreted in a Bayesian sense, as a MAP-estimate of logistic regression with a Generalised Gaussian Distribution (GGD) (known also as Generalised Laplacian) prior. Indeed, the regularisation term is, up to a constant, the log of independent zero-mean multivariate GGDs, $P(\mathbf{w}) = \prod_{j=1}^m GGD(w_j|0, \lambda, q)$ where

$$GGD(w_j|\mu_j, \lambda, q) = \frac{q\lambda^{1/q}}{2\Gamma(1/q)} \exp\{-\lambda|w_j - \mu_j|^q\}$$

We are interested in the case $q \in (0, 1]$, so the log prior is non-differentiable at zero, resulting in sparse estimates.

Using this interpretation, we now employ the PAC-Bayes methodology [4, 3] to analyse the L_q -regularised sparse classifier. Unlike in these works, our model priors are of course non-Gaussian.

PAC-Bayes analysis. The general statement of the PAC-Bayes theorem is the following (see e.g. [3]).

For all i.i.d. distributions D over the data, $\forall P(h)$ prior over the classifiers $h: \mathbf{X} \rightarrow Y$, and $\forall \delta \in (0, 1]$ risk,

$$\Pr_{S \sim D^n} \left\{ \forall Q(h): KL_+[Q_S||Q_D] \leq \frac{1}{n} [KL(Q||P) + \ln \frac{n+1}{\delta}] \right\} \geq 1-\delta \quad (2)$$

where $Q_D \equiv Pr_{(\mathbf{x}, y) \sim D} \{h(\mathbf{x}) \neq y\}$ is the true (generalisation) error and $KL_+[Q_S||Q_D] = \hat{Q}_S \ln \frac{\hat{Q}_S}{Q_D} + (1 - \hat{Q}_S) \ln \frac{1 - \hat{Q}_S}{1 - Q_D}$ for $Q_D \geq \hat{Q}_S$ (and 0 otherwise). Further, $\hat{Q}_S \equiv E_{h \sim Q} [\frac{1}{n} \sum_{i=1}^n I(h(\mathbf{x}_i) \neq y_i)]$ is the expected training error, and $KL(Q||P) = E_{h \sim Q} \ln \frac{Q(h)}{P(h)}$. The latter two quantities are detailed in the sequel specifically for our L_q -sparse classifier. We take Q of the form $GGD(w_j|\hat{w}_j, \tilde{\lambda}, \tilde{q})$, and $\tilde{\lambda}, \tilde{q}$ are tuned to tighten the bound.

The \hat{Q}_S term. One can obtain an upper bound on \hat{Q}_S by taking a set of samples from $Q(\mathbf{w})$ to estimate the empirical error on each training point, and using binomial tail

inversion (using up part of the specified δ) to obtain an upper bound [3]. A computationally cheaper alternative is to develop an analytic Gaussian approximation motivated by the central limit theorem. For the case $q = 1$ (Laplace priors), a similar approach was taken in [1], which we extend to $q \leq 1$, yielding the following:

$$\hat{Q}_S \approx \frac{1}{n} \sum_{i=1}^n \Phi \left\{ -\frac{y_i \hat{\mathbf{w}}^T \mathbf{x}_i}{\tilde{\lambda}^{-1/\tilde{q}} \sqrt{\frac{\Gamma(3/\tilde{q})}{\Gamma(1/\tilde{q})}} \|\mathbf{x}_i\|_2} \right\} \quad (3)$$

where $\Phi(\cdot)$ is the standard Gaussian cdf. Empirically, we found this approximation fairly accurate.

The $KL[Q||P]$ term. Due to the independent priors, we have $KL[Q||P] = \sum_{j=1}^m KL[Q(w_j)||P(w_j)]$. For GGDs with $q < 1$, unfortunately, this integral becomes analytically intractable in general. However we obtain an analytic upper bound, using the q -triangle inequality (valid for $q < 1$), which gives: $KL[Q(w_j)||P(w_j)] \leq$

$$\begin{aligned} &\leq KL[GGD(w_j|0, \tilde{\lambda}, \tilde{q})||GGD(w_j|0, \lambda, q)] + \lambda |\hat{w}_j|^q \quad (4) \\ &= \log \frac{\tilde{q} \tilde{\lambda}^{1/\tilde{q}} \Gamma(1/q)}{q \lambda^{1/q} \Gamma(1/\tilde{q})} + \frac{\lambda}{\tilde{\lambda}^{q/\tilde{q}}} \frac{\Gamma((q+1)/\tilde{q})}{\Gamma(1/\tilde{q})} - \frac{1}{\tilde{q}} + \lambda |\hat{w}_j|^q \end{aligned}$$

Observe that only the KL-terms of the non-zero estimates are affected by this approximation, since for components with $\hat{w}_j = 0$, the above inequality is satisfied with equality.

It is also indicative to look at the expression (4) in the case of low expected training error \hat{Q}_S . In this case it is known [4] that \hat{Q}_D is roughly proportional to $KL(Q||P)$. The minimum of (4) is achieved by setting $\tilde{\lambda} = \lambda$ and $\tilde{q} = q$, which gives the simple and self-explanatory expression:

$$\min_{\tilde{\lambda}, \tilde{q}} KL[Q||P] \leq \lambda \|\hat{\mathbf{w}}\|_q^q \quad (5)$$

We see in the case of low \hat{Q}_S , the components with zero-estimates incur no cost, which is nice, considering that we are dealing with a sparse classifier.

However, in general, the bound on Q_D can be tightened by taking both the KL-term and \hat{Q}_S into account when optimising for $\tilde{\lambda}$ and \tilde{q} .

References

- [1] B. Krishnapuram, et al. Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. IEEE PAMI, vol. 27, no. 6, pp. 957-968, 2005.
- [2] A. Kaban and R.J. Durrant. Learning with $L_{q < 1}$ vs. L_1 regularization in exponentially many irrelevant features. Proc. ECML'08; & ICML/UAI/COLT'08 Workshop.
- [3] J. Langford. Tutorial on practical prediction theory for classification. JMLR, 2005.
- [4] M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian Process classification. JMLR, 2003.
- [5] A.Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. Proc. ICML 2004.