Active learning for constrained regression using kernel beta regression models

Luis Montesano and Manuel Lopes

May 1, 2009

Abstract

In this poster we study active learning for supervised regression algorithms. We focus on a particular problem where the regression is constrained to an interval. This is the case, for instance, when the objective is to estimate the probability of success of a certain event given a set of input points. The regression problem is, thus, constrained to the zero-one interval and the training examples are also discrete, e.g. success or failure.

The constrained regression problem is usually solved using logistic regression [1]. This is in essence a generalized linear model for discrete outputs (e.g. classes) based on the logits of the probabilities. Active strategies for this type of techniques have been studied in [2] and pointed out a trade off between computational cost for experimental design based techniques and robustness issues that appear in more heuristic criteria. A support vector machine version of the logistic regression, the kernel logistic regression [3] uses the log-likelihood of the binomial distribution as the loss function and provides directly estimates of the probability. Constrained regression has also been studied using Beta models in economics [4], to model rates and proportions [5] and in psychological studies to account for skew and heteroscedasticity [6]. In this case, a parametric model is fitted maximizing the likelihood function using, for instance, Newton-Raphson or Fisher scoring. Better experimental results have been reported using alternative residuals in [7]. Bayesian versions of this regressors have been recently proposed based on a hierarchical model and priors on the parameters [8].

This poster proposes a new algorithm for such a constrained regression problem specially suited for the active learning framework. As logisitc regression, the algorithm uses a Binomial likelihood model. At each point of the input space, a conjugate Binomial-Beta model provides the distribution of the probability of success. The two parameters of the Beta at a particular point $\mathbf{x}_*$ are computed by accumulating evidence of successful and failed events at training points $\mathbf{x}_i$ using a kernel $K(\mathbf{x}_*, \mathbf{x}_i)$ The expression for the Beta distribution is

$$p(p_* \mid \mathbf{x}_*, \mathbf{X}_n, \mathbf{Y_n}) \quad \propto \prod_{i=0}^{n} \mathrm{Bin}(S_{*i}; p_*, S_{*i} + U_{*i}) \mathrm{Be}(p_*; \alpha_0, \beta_0) \quad = \mathrm{Be}\left(p_*; \sum_{i=1}^{n} S_{*i} + \alpha_0, \sum_{i=1}^{n} U_{*i} + \beta_0\right),$$

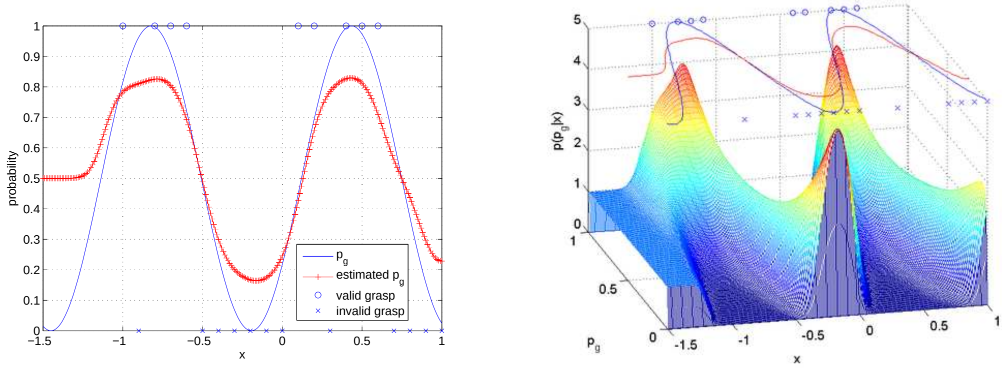
where $S_{*i} = K(\mathbf{x}_*, \mathbf{x}_i) S_i$ and $U_{*i} = K(\mathbf{x}_*, \mathbf{x}_i) U_i$ are the accumulated virtual number of successful and failed trials at point $\mathbf{x}_*$ given the successes (S_i) and failures (U_i) at training points $\mathbf{x}_i$, $i \in 1..n$. Figure 1 (a) shows the mean regression for a one dimensional problem and Fig. 1 (b) shows the predicted Beta distributions at each point of the input space.

Since we recover a full Beta distribution, an active learning strategy exploits this extra information, such as the variance (or other measurements of information), to select new training points in a pool based situation.The general active learning selection equation we will use is

$$\mathbf{x}_n \quad = \quad \arg\max_{\mathbf{x}_i \in \mathbf{X}_o} I(\mathbf{x}_i), \tag{1}$$

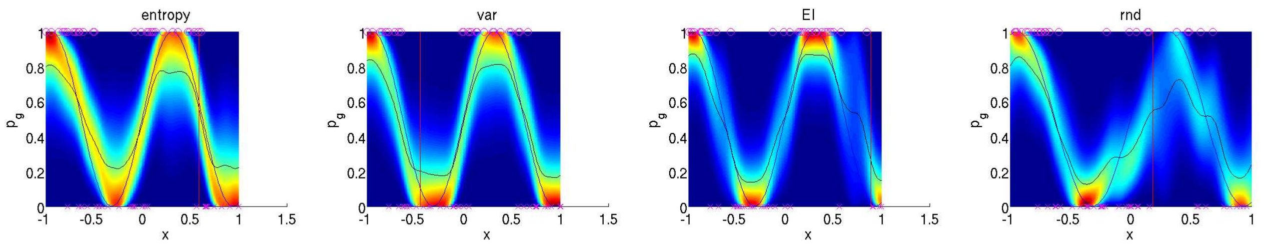
where $I(\mathbf{x})$ is a measure of the *improvement* in the regression after trying the point $\mathbf{x}$ from the set of all possible points $\mathbf{X}_o$. Such a strategy allows a great reduction in learning time/samples necessary to converge to a good approximation of the function. In addition to this, we also design an active criterium that focuses exploration on promising areas of the space, e.g. areas with an expected high probability of success. In this case, the function $I(\mathbf{x}_i)$ trades off exploration and exploitation for instance using the predicted mean $\bar{p}_i$ and variance $Var(p_i)$ of the candidates, $I(\mathbf{x}_i) = \bar{p}_i Var(p_i)$..

We provide results of our method on a simple simulated one dimensional problem (see Fig. 2 for a comparison of different active strategies) and on an higher dimensional one that consist on learning the probability of grasping points of an object [9]. In the latter, the robot has to actively try to grasp objects at different positions. Due to noise and partial information, the result of grasping is not deterministic. Thus, we learn the probability of grasping given the input features. We present results using a simulated dataset (with deterministic outputs) [9] and with real data acquired with a humanoid robot where the output of a grasp trial is actually non deterministic.



(a) (b)

Figure 1: Approximating a sinus varying p in a one dimensional input space. (a) Estimated mean. The 0-1 blue points are the observations generated from a Bernoulli using the true p (blue line). Failures are represented by crosses and successes by circles. The red line with marks is the approximated mean computed from the posterior. (b) Predicted posterior beta distributions for each point along x.



(a) (b) (c) (d)

Figure 2: The estimated parameter p after 60 queries. The points were selected based on (a) entropy, (b) variance, (c) EI and (d) randomly.

References

[1] J. Hilbe, *Logistic Regression Models*. Chapman & Hall/CRC Press, 2009.

[2] A. Schein and L. H. Ungar, "Active learning for logistic regression: an evaluation," *Machine Learning*, vol. 68, pp. 235–265, 2007.

[3] T. Hastie and T. Tibshirani, *Generalized Additive Models*. Chapman & Hall, 1990.

[4] P. Paolino, "Maximum likelihood estimation for models with beta distributed dependent variables," *Political Analysis*, vol. 9, pp. 325–346, 2001.

[5] S. L. P. Ferrari and F. Cribari-Neto, "Beta regression for modeling rates and proportions," *J. Applied Statistics*, vol. 31, pp. 799–815, 2004.

[6] M. Smithson and J. Verkuilen, "A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables," *Psychological Methods*, vol. 11, pp. 54–71, 2006.

[7] L. Espinheira, S. Ferrari, and F. Cribari-Neto, "On beta regression residuals," *Journal of Applied Statistics*, vol. 35, pp. 407–419, 2008.

[8] T. C. Martins and C. A. Ribeiro, "The use of several link functions on a beta regression model: a bayesian approach," in *Proceedings of the 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2008.

[9] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *International Journal of Robotics Research (IJRR)*, 2008.