Schedae Informaticae Vol. 25 (2016): 61–72
doi: 10.4467/20838476SI.16.005.6186

tfml 2017
theoretical foundations
of machine learning, Kraków

Optimization of ℓ^p-regularized Linear Models via Coordinate Descent

Jacek Klimaszewski, Marcin Korzeń
Group, Department
Faculty of Computer Science and Information Technology
ul. Żołnierska 49, 71-210, Szczecin, Poland
e-mail: {*jklimaszewski, mkorzen*}*@wi.zut.edu.pl*

Abstract. In this paper we demonstrate, how ℓ^p-regularized univariate quadratic loss function can be effectively optimized (for $0 \leqslant p \leqslant 1$) without approximation of penalty term and provide analytical solution for $p = \frac{1}{2}$. Next we adapt this approach for important multivariate cases like linear and logistic regressions, using Coordinate Descent algorithm. At the end we compare sample complexity of ℓ^1 with $\ell^p, 0 \leqslant p < 1$ regularized models for artificial and real datasets.

Keywords: Classification, Coordinate Descent, Regression, Sparsity

1. Introduction

In the supervised learning there are two (usually numeric) matrices: $\mathbf{X}_{n \times d}$ and $\mathbf{y}_{n \times 1}$, where n stands for number of observations and d represents number of attributes. The goal is to build such a model, that for unseen examples it would predict correct answers. Therefore final model should contain only relevant features, that were selected at training stage. One of the way to do this is to include regularization term in the loss function, which penalizes coefficients in the model.

Received: 11 December 2016 / Accepted: 30 December 2016

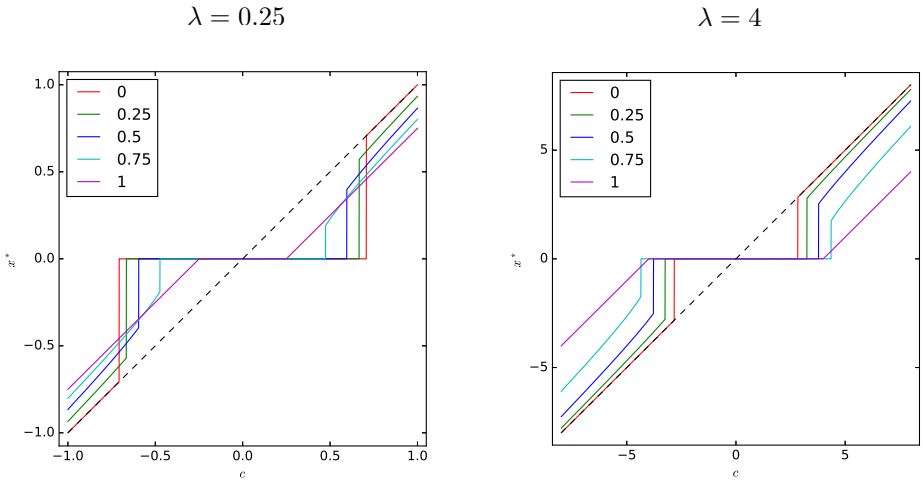

Figure 1. Plot of $x^* = \arg\min_x \frac{1}{2} \cdot (c - x)^2 + \lambda \cdot |x|^p$ for various exponent p and λ.

1.1. Related Work

Historically regularization was used for the first time to solve ill-posed problems [1]. The so called *Tikhonov regularization* (also known as *ridge regression*) penalizes squared ℓ^2-norm of coefficients. It is known that this type of regularization shrinks correlated features, but it also produces dense solutions (all coefficients in the model are non-zero). The same situation is observed for bridge regression (ℓ^p for $p \in (1, 2)$) [2], because regularization term is strictly convex.

Next, ℓ^1-norm was tried and it drew a big attention, because it produces sparse solutions [3, 4]. ℓ^1-regularized problem with convex loss function is still convex, but its derivative has discontinuity at origin, which causes certain coefficients to be set exactly to zero. In the Figure 1 it can be seen that solution is shifted from the true coefficient (dashed line), what results in biased models.

When $0 \leqslant p < 1$, ℓ^p-regularized problem with convex loss function becomes non-convex, making optimization more difficult due to many local minima. On the other hand, resulting model has smaller bias. Work by [5] treated non-convex regularization, but local quadratic approximation was used to approximate concavity — this approach is sensitive to initialization, as it can give different solutions for different initial points (once a coefficient is set to zero, it will stay at zero). ℓ^p-"norm" was also mentioned in [6], but authors abandoned using coordinate descent procedure in this case, because they encountered some instability (caused by discontinuity in path of solutions, what is depicted in the Figure 1) and impossibility of converging to the global minimum (using Multi-stage Local Linear Approximation), even for some univariate case. Some non-convex regularizers were also studied in [7] and ℓ^p quasi-norm was considered in [8].

2. The Algorithm for Univariate Case

Consider a function (also known as *proximal operator* [9]):

$$f(x) = \frac{\mu}{2} \cdot (c - x)^2 + \lambda \cdot |x|^p, \tag{1}$$

where $p \in [0, 1]$, $c \in \mathbb{R}$ is a true coefficient, $x \in \mathbb{R}$ is its estimate, $\mu > 0$ controls strength of squared error and $\lambda \geq 0$ controls strength of regularization. One may note that μ can be fused with λ and equation (1) can be written in a different way:

$$f(x) = \frac{1}{2} \cdot (c - x)^2 + \rho \cdot |x|^p, \text{ where } \rho = \frac{\lambda}{\mu}, \tag{2}$$

but this form was deliberately omitted for better clarification.

For $p = 1$ equation (1) is still convex and its unique global minimum is given by a soft-thresholding formula [10]:

$$x^* = \operatorname{sgn}(c) \cdot \max\left(0,\ |c| - \frac{\lambda}{\mu}\right). \tag{3}$$

The case $p = 0$ is known as hard-thresholding [10], as it minimizes number of non-zero coefficients. Minimum of equation (1) is either 0 or c[1].

When $p \in (0, 1)$, regularization term causes non-convexity, what can be seen in the Figure 2. In this case equation (1) may have 1 minimum (either c for $\lambda = 0$ or 0 for sufficiently large λ) or 2 minima.

2.1. Special Case of $p = \frac{1}{2}$

When $x \geqslant 0$, equation (1) can be transformed into quartic function using substitution $t = x^{\frac{1}{2}}$:

$$g(t) = \frac{\mu}{2} \cdot (c - t^2)^2 + \lambda \cdot t, \tag{4}$$

whose roots (and extrema) can be calculated analytically. Subtracting $\frac{\mu}{2} \cdot c^2$ from equation (4) yields:

$$t \cdot \left(\frac{\mu}{2} \cdot t^3 - \mu \cdot c \cdot t + \lambda\right) = 0. \tag{5}$$

Now we need to find such $\lambda_{\text{critical}}$ for which equation (5) has double root (that coincides with minimum). Using ideas presented in [11] it can be shown that:

$$\lambda_{\text{critical}} = \mu \cdot \left(\frac{2}{3} \cdot c\right)^{\frac{3}{2}}. \tag{6}$$

[1] If $f(\mathbf{x}) = \frac{\mu}{2} \cdot \|\mathbf{c} - \mathbf{x}\|_2^2 + \lambda \cdot \sum_j |x_j|^p$, then each x_j^* can be computed independently (x_j^* is either 0 or c_j). This does not hold in general case $f(\mathbf{x}) = \frac{\mu}{2} \cdot \|\mathbf{c} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \cdot \sum_j |x_j|^p$.

If $\lambda > \lambda_{\text{critical}}$, then equation (4) has global minimum at zero. Otherwise its stationary points need to be found. Differentiating with respect to t yields cubic equation:

$$t^3 - ct + \frac{\lambda}{2\mu} = 0. \tag{7}$$

In this case equation (7) has 3 real roots (see [11]):

$$\begin{cases} \alpha &= 2 \cdot \sqrt{\frac{c}{3}} \cdot \cos\theta, \\ \beta &= 2 \cdot \sqrt{\frac{c}{3}} \cdot \cos\left(\theta + \frac{2\pi}{3}\right), \\ \gamma &= 2 \cdot \sqrt{\frac{c}{3}} \cdot \cos\left(\theta + \frac{4\pi}{3}\right), \end{cases} \tag{8}$$

where

$$\theta = \frac{1}{3} \cdot \arccos\left(-\frac{\lambda}{4\mu\left(\frac{c}{3}\right)^{\frac{3}{2}}}\right). \tag{9}$$

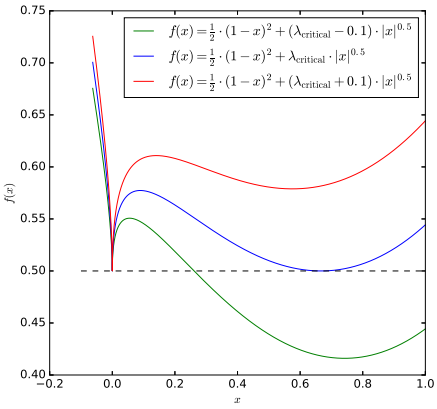

Figure 2. Plot of eq. (1) for different λ.

α is a local minimum of equation (4), γ is a local maximum and β is ignored, because it is negative. Since $t = x^{\frac{1}{2}}$, α need to be squared to obtain x^*. Negative case is handled similarly.

2.2. Algorithm for General Case of $p \in (0, 1)$

In general case Newton's method [12] can be used to find minimum of equation (1), because it is differentiable for $x \neq 0$:

$$f'(x) = \mu \cdot (x - c) + \lambda \cdot p \cdot \operatorname{sgn}(x) \cdot |x|^{p-1}, \tag{10}$$

$$f''(x) = \mu + \lambda \cdot p \cdot (p-1) \cdot |x|^{p-2}. \tag{11}$$

If $c = 0$, then $x = 0$ is a global minimum.

$\lambda_{\text{critical}}$ in general case was found in a similar way to case $p = \frac{1}{2}$ — formula $f(x) - \frac{\mu}{2}c^2$ has 2 minima that coincide with roots when $\lambda = \lambda_{\text{critical}}$. To find $\lambda_{\text{critical}}$, the following system of equations has to be solved:

$$\begin{cases} f(x) - \frac{\mu}{2}c^2 = 0 \\ f'(x) = 0 \end{cases} \tag{12}$$

It can be solved via substitution — the solution is:

$$\begin{cases} x = \frac{2-2p}{2-p} \cdot c, \\ \lambda_{\text{critical}} = \frac{\mu|c|}{2-p} \cdot \left(\frac{2-2p}{2-p} \cdot |c|\right)^{1-p}. \end{cases} \tag{13}$$

Now minimization of (1) is much easier: firstly, value $\lambda_{\text{critical}}$ is computed and then condition $\lambda_{\text{critical}} \leq \lambda$ is checked — if it is met, then there is nothing to do, because global minimum is at zero (we select $x = 0$ even when $\lambda_{\text{critical}} = \lambda$, though global minimum is not unique in that case). Otherwise Newton's method is launched for a starting point $x_0 = c$ (it is very close to the minimum, so only a few iterations are needed).

3. The algorithm for multivariate case

In this section it is shown how Coordinate Descent method is applied to minimize ℓ^p-regularized residual sum of squares (RSS). Then this approach is used to estimate coefficients of the logistic regression model via iteratively reweighted least squares.

3.1. Basic Algorithm for Linear Regression

Coefficients w_j of the linear regression model are estimated by minimization of residual sum of squares (RSS). ℓ^p-regularized loss function has a form (intercept w_0 is not regularized):

$$L(\mathbf{w}) = \frac{1}{2} \cdot \sum_i \left(y_i - w_0 - \sum_j x_{ij} w_j \right)^2 + \lambda \cdot \sum_j |w_j|^p, \tag{14}$$

where y_i is the i-th value of the variable to be predicted and $\mathbf{x}_i$ is the i-th row of the explanatory matrix $\mathbf{X}$. Differentiating equation (14) with respect to w_j yields following formulae:

$$\frac{\partial L}{\partial w_j} = \sum_i x_{ij}^2 \cdot \left(w_j - \frac{\sum_i x_{ij} \cdot \left(y_i - \sum_{k \neq j} x_{ik} w_k \right)}{\sum_i x_{ij}^2} \right) + \lambda \cdot p \cdot \operatorname{sgn}(w_j) \cdot |w_j|^{p-1}, \tag{15}$$

$$\frac{\partial^2 L}{\partial w_j^2} = \sum_i x_{ij}^2 + \lambda \cdot p \cdot (p-1) \cdot |w_j|^{p-2} \tag{16}$$

From equations (15) and (16) it can be seen that $c_j = \frac{\sum_i x_{ij} \cdot \left(y_i - \sum_{k \neq j} x_{ik} w_k\right)}{\sum_i x_{ij}^2}$ and $\mu_j = \sum_i x_{ij}^2$, so:

$$\frac{\partial L}{\partial w_j} = \mu_j \cdot (w_j - c_j) + \lambda \cdot p \cdot \operatorname{sgn}(w_j) \cdot |w_j|^{p-1}, \tag{17}$$

$$\frac{\partial^2 L}{\partial w_j^2} = \mu_j + \lambda \cdot p \cdot (p-1) \cdot |w_j|^{p-2}. \tag{18}$$

Now it is clear that equations (17) and (18) have the same form as equations (10) and (11) respectively. Hence Coordinate Descent procedure can be constructed:

Algorithm 1 Coordinate Descent procedure for ℓ^p-regularized Linear Regression.

```
while has not converged do
    w_0 ← (1/n) · Σ_i ( y_i − Σ_j w_j x_ij )
    for j ← 1 to d do
        Compute c and λ_critical for j-th coordinate.
        if λ_critical ≤ λ then
            w_j ← 0
        else
            w_j ← NewtonAlgorithm(∂L/∂w_j, ∂²L/∂w_j², c)
        end if
    end for
end while
```

The algorithm is stopped when it exceeds maximum number of iterations or maximum relative difference between $w_j^{(k)}$ and $w_j^{(k+1)}$ falls below $\epsilon = 10^{-5}$ (or some other value specified by user).

3.2. Basic Algorithm for Logistic Regression

A similar approach can be adapted for fitting the logistic regression model. Here we consider the case where $\mathbf{y}$ is a binary variable ($y_i \in \{0, 1\}$):

$$P(y_i = 1 | \mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp(-w_0 - \sum_j w_j x_{ij})}, \tag{19}$$

$$P(y_i = 0 | \mathbf{x}_i, \mathbf{w}) = 1 - P(y_i = 1 | \mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp(w_0 + \sum_j w_j x_{ij})}. \tag{20}$$

A common way of estimation of model's coefficients is minimization of the negative log-likelihood function (or, equivalently, maximization of the likelihood function). The ℓ^p-regularized negative log-likelihood function has a form:

$$L(\mathbf{w}) = -\sum_i y_i \cdot \log\left(p\left(\mathbf{x}_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(\mathbf{x}_i\right)\right) + \lambda \cdot \sum_j |w_j|^p, \tag{21}$$

where $p(\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i, \mathbf{w})$.

Equation (21) does not have closed-form solution, but it can be locally approximated by quadratic function [13]. Introduction of weights:

$$\alpha_i = p(\mathbf{x}_i) \cdot \left(1 - p(\mathbf{x}_i)\right), \tag{22}$$

$$z_i = w_0 + \sum_j w_j x_{ij} + \frac{y_i - p(\mathbf{x}_i)}{\alpha_i}, \tag{23}$$

allows us to rewrite equation (21) to the form:

$$L(\mathbf{w}) = \frac{1}{2} \cdot \sum_i \alpha_i \cdot \left(z_i - w_0 - \sum_j w_j x_{ij} \right)^2 + \lambda \cdot \sum_{j=1} |w_j|^p. \tag{24}$$

Now minimization of equation (21) is changed to the successive minimization of equation (24) — for $\mathbf{w}^{(k)}$ update weights and compute $\mathbf{w}^{(k+1)}$ until convergence.

Finally a similar algorithm to the Algorithm 1 of fitting logistic regression model is presented below:

Algorithm 2 Coordinate Descent for ℓ^p-regularized Logistic Regression.

while has not converged **do**
  Update α and $\mathbf{z}$ using current $\mathbf{w}$.
  while not converged **do**
    $w_0 \leftarrow \frac{\sum_i \alpha_i \cdot \left(z_i - \sum_j w_j x_{ij}\right)}{\sum_i \alpha_i}$
    for $j \leftarrow 1$ **to** d **do**
      Compute c and $\lambda_{\text{critical}}$ for j-th coordinate.
      if $\lambda_{\text{critical}} \leq \lambda$ **then**
        $w_j \leftarrow 0$
      else
        $w_j \leftarrow \text{NewtonAlgorithm}(\frac{\partial L}{\partial w_j},\ \frac{\partial^2 L}{\partial w_j^2},\ c)$
      end if
    end for
  end while
end while

3.3. Improvements to the Basic Algorithms

Ad hoc implementation of above algorithms is not efficient for large datasets. To improve it, ideas described in [14] were used — i.e. naive updates, pathwise coordinate descent and computation over *active set* of features.

4. Experiments

Experimental part was written in Python. Procedures for estimation of coefficients of ℓ^p-regularized linear and logistic regression were implemented in C++ and called from Python script via *ctypes* package. Vectorized matrices using column-major order were passed to the C++ routines to speed-up computation. Some parts of *glmnet* [14] were used during implementation of our solution. In each experiment tolerance ϵ was set to 10^{-5}.

We use three test data sets:

1. DataSet#1 is artificial set of size $100 \times 1000(4)$ (consisting of 100 samples with 1000 attributes drawn from multivariate normal distribution and only 4 significant attributes), output is a linear combination of 4 significant variables (in case of logistic regression it was sign of this linear combination);
2. DataSet#2 is artifical set of size $100 \times 1000(32)$ generated similarly;
3. DataSet#3 is a Golub's `Leukemia` dataset [15], preprocessed by [16]. This dataset has 38 training samples and 34 test samples.

4.1. Impact of p on coefficients' paths

The first experiment shows the coefficients' path for linear and logistic regression models for $p = 0, 0.333, 0.667, 1$. The results are presented in the Figures 3 and 4. The vertical dashed line shows value of λ that yields optimal model (with the highest accuracy). The accuracy measure in the case of linear regression is RSS, and for logistic regression it is the accuracy of classification. As one can see, in both cases ℓ^1-regularized regressions include some number of random attributes in the optimal model. In the case $p < 1$ the optimal model selects 4 significant attributes correctly — this shows a qualitative difference between ℓ^1 and $\ell^p, p < 1$ penalty terms. The second observation is that for smaller exponents p coefficients' path is more robust with respect to λ (for $p = 0$ coefficients' paths are piecewise constant).

It can also be noted that in case $d \gg n$ when λ is close to 0, coefficients of logistic regression wander off to $\pm\infty$ in order to achieve probabilities of 0 or 1, but this is natural.

4.2. Sample complexity of Logistic Regression

In the next experiment we compared influence of exponent p on sample complexity of the model. We varied number of samples in the train set, next we tested each classifier using test set, we took coefficients from the solution's path (computed for 65 values

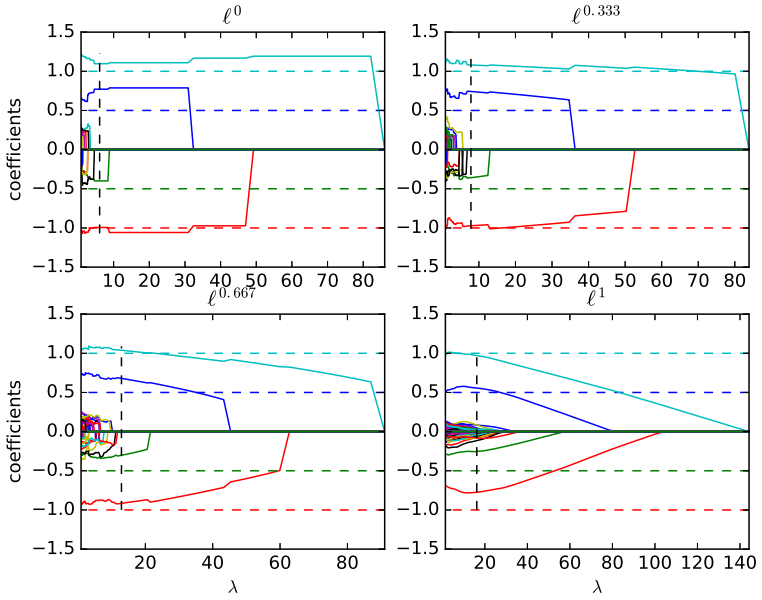

Figure 3. Paths of coefficients for the linear regression for DataSet#1; horizontal dashed lines represent true coefficients, vertical dashed line depicts value of λ for which model has the lowest RSS, tested via 10-fold cross validation procedure.

of λ, $\lambda_{\min} = 0.01 \cdot \lambda_{\max}$) and we computed prediction for each solution on the path. This process was repeated 50 times and the accuracy was averaged. Best results for each sample size are presented in the Figure 5 for DataSet#1 and in the Figure 6 for DataSet#2. As the number of training samples rises, accuracies of all models grows. It can be seen that accuracies of ℓ^1- and $\ell^{0.75}$-regularized models are roughly the same in the beginning, but later $\ell^{0.75}$ is superior to ℓ^1. Surprisingly, models for $p \leqslant 0.5$ gave almost the same accuracies. Generally results show that models with smaller exponent p achieve the desired accuracy earlier.

4.3. Test on Leukemia dataset

Logistic regression model was trained on preprocessed `Leukemia` dataset using leave-one-out cross-validation for 100 values of λ and $\lambda_{\min} = 0.001 \cdot \lambda_{\max}$. Table 1 presents results of the experiment.

It can be seen that all models for $p < 1$ gave sparser solution than ℓ^1-regularized model. Also models for $p \geqslant 0.5$ are more accurate (higher area under ROC curve). Although $\ell^{0.5}$-regularized model seems to be superior, it is hard to select the best model in this case, because dataset is relatively small.

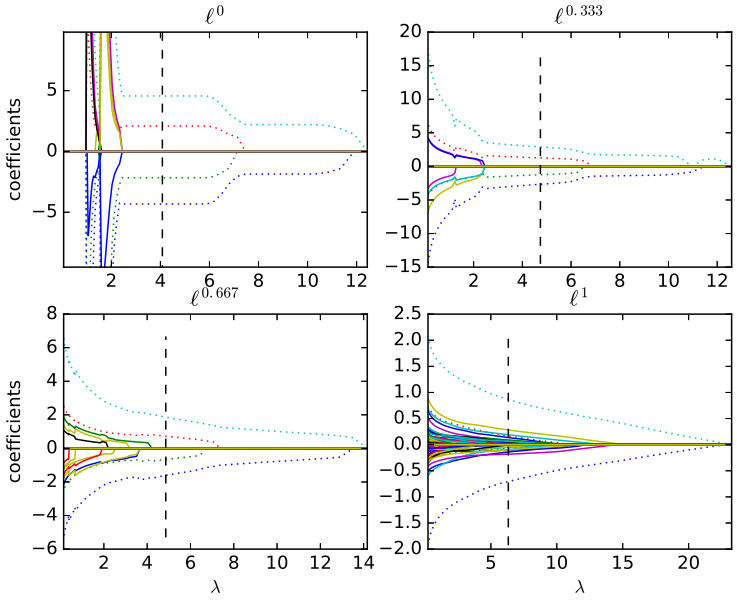

Figure 4. Paths of coefficients for the logistic regression for DataSet#1; vertical dashed line depicts value of λ for which model has the highest classification accuracy, tested via 10-fold cross-validation procedure; dotted line corresponds to relevant features.

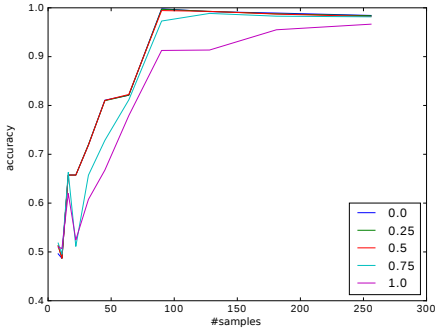

Figure 5. Results of the experiment with 4 relevant features.

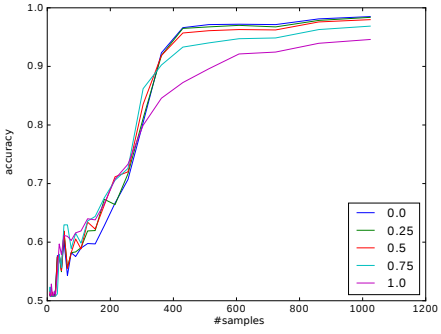

Figure 6. Results of the experiment with 32 relevant features.

5. Conclusions

In the paper we have shown that ℓ^p-regularized regressions can be effectively fitted via adapted version of pathwise coordinate descent algorithm. The results show that in some cases models with penalty function for smaller exponent p can lead to models with some attractive features like:

Table 1. Accuracy on the test set.

p	Timing	Accuracy	AUC	#nnz
0.0	1.78	32/34	0.936	2
0.25	16.17	30/34	0.925	1
0.5	19.20	33/34	0.993	1
0.75	15.04	32/34	0.968	2
1.0	7.20	31/34	0.989	7

- sparser solutions;
- more robust coefficients' paths with respect to λ;
- smaller exponents p, close to $p = 0$, have a better selecting properties in the case of independent attributes.

The case of correlated variables, which was in fact omitted in the paper, needs further research, because all considered models should be corrected in the similar way to elastic-net model [17].

6. References

[1] Tikhonov A.N., *On the stability of inverse problems (in Russian).* Doklady Akademii Nauk SSSR, 1943, 39 (5), pp. 195–198.

[2] Frank I.E., Friedman J.H., *A Statistical View of Some Chemometrics Regression Tools.* Technometrics, 1993, 35, pp. 109–148.

[3] Williams P.M., *Bayesian Regularisation and Pruning using a Laplace Prior.* Neural Computation, 1994, 7, pp. 117–143.

[4] Tibshirani R., *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society, Series B, 1996, 58, pp. 267–288.

[5] Fan J., Li R., *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.* Journal of the American Statistical Association, 2001, 96, pp. 1348–1360.

[6] Mazumder R., Friedman J., Hastie T., *SparseNet: Coordinate Descent With Nonconvex Penalties.* Journal of the American Statistical Association, 2011, 106 (495), pp. 1125–1138.

[7] Nikolova M., *Analysis of the Recovery of Edges in Images and Signals by Minimizing Nonconvex Regularized Least-Squares.* Multiscale Modeling & Simulation, 2005, 4 (3), pp. 960–991.

[8] Bredies K., Lorenz D., Reiterer S., *Minimization of Non-smooth, Non-convex Functionals by Iterative Thresholding.* Journal of Optimization Theory and Applications, 2015, 165, pp. 78–112.

[9] Moreau J.J., *Fonctions convexes duales et points proximaux dans un espace hilbertien.* Comptes Rendus de l'Académie des Sciences (Paris), Série A, 1962, 255, pp. 2897–2899.

[10] Donoho D., Johnstone I., *Ideal Spatial Adaptation by Wavelet Shrinkage.* Biometrika, 1994, 81, pp. 425–455.

[11] Nickalls R.W.D., *A New Approach to Solving the Cubic: Cardan's Solution Revealed.* The Mathematical Gazette, 1993, 77 (480), pp. 354–359.

[12] Kincaid D., Cheney W., *Numerical Analysis: Mathematics of Scientific Computing.* American Mathematical Society, 2002.

[13] Green P.J., *Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives.* Journal of the Royal Statistical Society. Series B (Methodological), 1984, 46 (2).

[14] Friedman J., Hastie T., Tibshirani R., *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010, 33 (1), pp. 1–22.

[15] Golub T., Slonim D., Tamayo P., Huard C., Gaasenbeek M., Mesirov J., Coller H., Loh M., Downing J., Caligiuri M., Bloomfield C., Lander E., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999, 286 (5439), pp. 531–537.

[16] Dettling M., *BagBoosting for Tumor Classification with Gene Expression Data.* Bioinformatics, 2004, 20 (18), pp. 3583–3593.

[17] Zou H., Hastie T., *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society, Series B, 2005, 67, pp. 301–320.