Feature selection, L1 vs. L2 regularization, and rotational invariance

Andrew Ng
ICML 2004

Presented by Paul Hammon
April 14, 2005

Outline

1. Background information
2. L_1-regularized logistic regression
3. Rotational invariance and L_2-regularized logistic regression
4. Experimental setup and results

Overview

The author discusses regularization as a feature selection approach.

For logistic regression he proves that L_1-based regularization is superior to L_2 when there are many features.

He proves lower bounds for the sample complexity: the number of training examples needed to learn a classifier.

3

L_1 vs. L_2 regularization

Sample complexity of L_1-regularized logistic regression is logarithmic in the number of features.

The sample complexity of L_2-regularized logistic regression is linear in the number of features.

Simple experiments verify the superiority of L_1 over L_2 regularization.

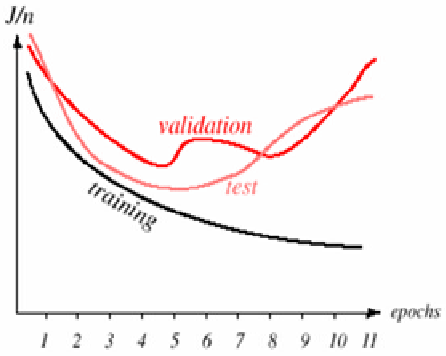
4

Background: Overfitting

Supervised learning algorithms often over-fit training data.

Overfitting occurs when there are so many free parameters that the learning algorithm can fit the training data too closely.

This increases the generalization error.



(Duda, Hart, & Stork 2001)

The degree of overfitting depends on several factors:
- Number of training examples—more are better
- Dimensionality of the data—lower dimensionality is better
- Design of the learning algorithm—regularization is better

5

Overfitting example

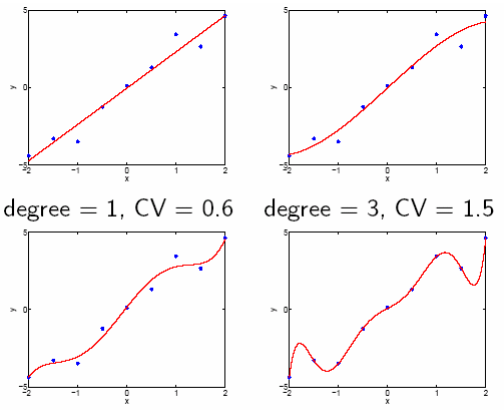
Consider the learning algorithm called polynomial regression.

Polynomial regression allows one to find the best fit polynomial to a set of data points.

Leave-one-out cross-validation estimates how well a learning algorithm generalizes

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - f(x^{(i)}; q^{(n-i)})^2$$

where $y^{(i)}$ is the class label,
$x^{(i)}$ is the training example,
f() is the classifier,
$?^{(n-i)}$ is the parameters trained without the i^{th} sample.



(from T. Jaakkola lecture notes)

6

VC-dimension

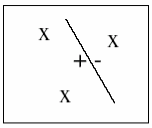
Vapnik-Chervonenkis (VC) dimension measures the geometric complexity of a classifier.

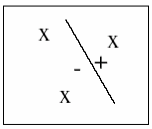
VC-dimension equals to the number of points the classifier can "shatter."

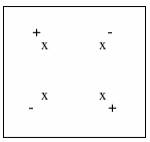
A classifier shatters a set of points by generating all possible labelings.

For example, the VC-dimension of 2-D linear boundaries is 3.

The VC-dimension for most models grows roughly linearly in the number of model parameters (Vapnik, 1982).







(from T. Jaakkola lecture notes)

7

Sample complexity

Recall that sample complexity is number of training examples needed to learn a classifier.

For (unregularized) discriminative models, the sample complexity grows linearly with VC-dimension.

As the number of model parameters increases, more training examples are necessary to generalize well.

This is a problem for learning with small data sets with large numbers of dimensions.

8

Regularization and model complexity

Adding regularization to a learning algorithm avoids overfitting.

Regularization penalizes the complexity of a learning model.

Sparseness is one way to measure complexity. Sparse parameter vectors have few non-zero entries

Regularization based on the zero-norm maximizes sparseness, but zero-norm minimization is an NP-hard problem (Weston et al. 2003).

Regularization based on the L_1 norm drives many parameters to zero. L_2 norm regularization does not achieve the same level of sparseness (Hastie et al 2001).

$$\| q \|_0 = \sum_{i=1}^{n} | q_i |^0 = \text{sum of non-zero entries}$$

$$\| q \|_1 = \sum_{i=1}^{n} | q_i |^1$$

$$\| q \|_2 = \left(\sum_{i=1}^{n} q_i^2 \right)^{1/2}$$

9

Logistic regression

Logistic regression (LR) is a binary classifier.

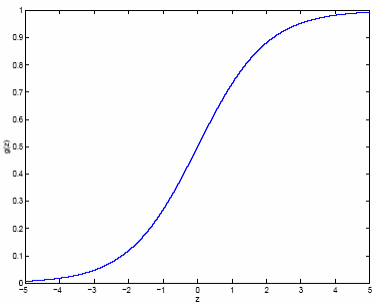
The LR model is

$$p(y = 1 | x; q) = \frac{1}{1 + \exp(-q^T x)} \qquad (1)$$

where y = {0, 1} is the class label,
x is a training point,
? is the parameter vector,
and the data is assumed to be drawn i.i.d. from a distribution *D*.

A logistic function turns linear predictions into [0, 1].

To simplify notation, each point *x* is formed by adding a constant feature, $x = [x_0, 1]^T$. This removes the need for a separate offset term.



The logistic function

$$g(z) = \frac{1}{1 + \exp(-z)}$$

(from A. Ng lecture notes)

10

Training regularized LR

To learn LR parameters we use maximum likelihood.

Using the model in (1) and an i.i.d. assumption, we have

$$\log\text{-likelihood}\quad \ell(q) = \log \prod_i p(y^{(i)} \mid x^{(i)}; q) = \sum_i \log p(y^{(i)} \mid x^{(i)}; q)$$

where i indexes the training points.

We maximize the regularized log-likelihood

$$\hat{q} = \arg\max_q \ell(q) - aR(q) \qquad (2)$$

$$\text{with } R(q) \in \{\| q \|_1, \| q \|_2^2\}$$

where a determines the amount of regularization.

Training regularized LR

An equivalent formulation to (2) is

$$\max_q \sum_i \log p(y^{(i)} \mid x^{(i)}; q) \qquad (3)$$

$$\text{s.t.} \quad R(q) \le B$$

For every a in (2), and equivalent B in (3) can be found.

This relationship can be seen by forming the Lagrangian.

This formulation has the advantage that the regularization parameter B bounds the regularization term R(?).

The algorithm and proof for L1-regularized LR use (3).

Metrics of generalization error

One metric for error is negative log-likelihood (log-loss)

$$e_l(q) = E_{(x,y)\sim D}[-\log p(y \mid x;q)]$$

where the subscript (x,y)~D indicates that the expectation is for test samples drawn from D.

This has a corresponding empirical log-loss of

$$\hat{e}_l(q) = \frac{1}{m}\sum_{i=1}^{m} -\log p(y^{(i)} \mid x^{(i)};q)$$

The proofs in this paper use this log-loss function.

Another metric is the misclassification error:

$$d_m(q) = P_{(x,y)\sim D}[t(g(q^T x)) \neq y]$$

where t(z) is a threshold function equal to 0 for z < 0.5 and equal to 1 otherwise. This also has a corresponding empirical log-loss of $\hat{d}_m$.

13

Regularized regression algorithm

The L_1-regularized logistic regression algorithm is:

1. Split the training data S into training set S_1 for the first (1-?)m examples, and hold-out set S_2 with the remaining ?m examples.

2. For B = 0, 1, 2, …, C,
   Fit a logistic regression model to the training set S_1 using

   $$\max_q \sum_i \log p(y^{(i)} \mid x^{(i)};q)$$

   $$\text{s.t.} \quad R(q) \leq B$$

   Store the resulting parameter vector as $?_B$.

3. Pick the $?_B$ that produces the lowest error score on the hold-out set S_2.

The theoretical results in this paper only apply to the log-loss error e_l.

14

L_1-regularized logistic regression

Theorem 1:

Let any e > 0, d > 0, K = 1, and
let m be the number of training examples and n be the number of features, and C = rK (the largest value of B tested).

Suppose there exists a parameter vector ?* such that
(a) only r components of ?* are non-zero, and
(b) every component of ?* = K

We want the parameter output by our learning algorithm to perform nearly as well as ?*:

$$e_l(\hat{q}) \le e_l(q^*) + e$$

To guarantee this with probability at least 1-d, it suffices that

$$m = \Omega((\log n) \cdot poly(r, K, \log(1/d), 1/e, C))$$

where O is a lower bound: O(g(n))=f(n) is defined by saying for some constant c>0 and large enough n, f(n)=c g(n).

15

Background of theorem 1 proof

The proof of this theorem relies on covering number bounds (Anthony & Bartlett, 1999), the detail of which is beyond the scope of this talk.

Covering numbers are used to measure the size of a parametric function family (Zhang, 2002).

More details can be found in (Anthony & Bartlett, 1999).

16

Discussion of theorem 1

Logistic regression using the L_1 norm has a sample complexity that grows logarithmically with the number of features.

Therefore this approach can be effectively applied when there are many more irrelevant features than there are training examples.

This approach can be applied to L_1 regularization of generalized linear models (GLMs).

GLM motivation: logistic regression

For logistic regression we have

$$p(y=1 \mid x;\theta) = \frac{1}{1+\exp(-\theta^T x)} = g(\theta^T x)$$

The distribution is Bernoulli with

$$p(y=0 \mid x;\theta) = 1 - p(y=1 \mid x;\theta)$$

The conditional expectation is

$$E[y \mid x] = 1\,p(y=1 \mid x;\theta) + 0\,p(y=0 \mid x;\theta) = g(\theta^T x)$$

To summarize, for logistic regression,

$$E[y \mid x] = g(\theta^T x)$$

$$p(y \mid x) \text{ is a Bernoulli distribution}$$

GLM motivation: linear regression

For linear regression we have

$$p(y \mid x; q, s^2) = N(q^T x, s^2) = \frac{1}{(2p)^{1/2} s} e^{-\frac{1}{2s^2}(y - q^T x)^2}$$

The conditional expectation is

$$E[y \mid x] = q^T x$$

To summarize, for linear regression,

$$E[y \mid x] = q^T x$$

$p(y \mid x)$ is a Gaussian distribution

19

GLMs

GLMs are a class of generative probabilistic models which generalize the setup for logistic regression (and other similar models) (McCullagh & Nelder, 1989).

The generalized linear model requires:

1. The data vector x enters the model as a linear combination with the parameters, $?^T x$.

2. The conditional mean μ is a function $f(?^T x)$ called the response function

3. The observed value y is modeled as distribution from an exponential family with conditional mean μ.

For logistic regression,
- $f(?^T x)$ is the logistic function $g(z) = 1/(1 + \exp(-z))$
- $p(y|x)$ is modeled as a Bernoulli random variable

20

The exponential family

GLMs involve distributions from the exponential family

$$p(x \mid h) = \frac{1}{Z(h)} h(x) \exp\{h^T T(x)\}$$

where ? is known as the natural parameter,
h(x) and Z(?) are normalizing factors, and
T(X) is a sufficient statistic

The exponential family includes many common distributions such as
- Gaussian
- Poisson
- Bernoulli
- Multinomial

Rotational invariance

For x in R^n and rotation matrix M, Mx is x rotated about the origin by some angle.

Let $M_R = \{M \text{ in } R^{nxn} | MM^T = M^TM = I, |M| = +1\}$ be the set of rotation matrices.

For a training set $S = \{(x_i, y_i)\}$, MS is the training set with inputs rotated by M.

Let L[S](x) be a classifier trained using S.

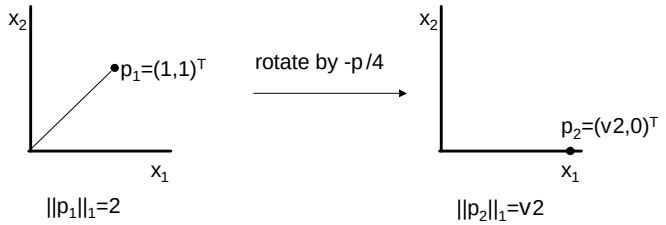
A learning algorithm L is **rotationally invariant** if L[S](x) = L[MS](Mx).

Rotational invariance and the L_1 norm

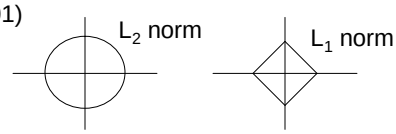
Regularized L_1 regression is not rotationally invariant.

The regularization term $R(?) = ||?||_1 = ?_i|?_i|$ causes this lack of invariance.

Observation: Consider the L_1 norm of a point and a rotated version of that point.



Contours of constant distance show circular symmetry for the L_2 but not the L_1 norm. (Hastie et al,2001)



23

Rotational invariance and L_2 regularization

Proposition: L_2-regularized logistic regression is rotationally invariant.

Proof:

Let S, M, x be given and let S' = MS and x' = Mx, and recall that $M^TM=MM^T= I$.

Then $$\frac{1}{1+\exp(-q^T x)} = \frac{1}{1+\exp(-(q^T(M^TM)x))} = \frac{1}{1+\exp(-(Mq)^T(Mx))}$$

so $$p(y \mid x;q) = p(y \mid Mx;Mq)$$

Recall that regularized log-likelihood is

$$J(q) = \sum_i \log p(y^{(i)} \mid x^{(i)};q) - aR(q)$$

Also note that regularization term $R(q) = q^Tq = q^T(M^TM)q = (Mq)^T(Mq) = R(Mq)$

Define $$J'(q) = \sum_i \log p(y^{(i)} \mid Mx^{(i)};Mq) - aR(Mq)$$

Clearly J(?) = J'(M?), and .

24

Rotational invariance and L2 regularization

Let $\hat{q} = \arg\max_q J(q)$ be the parameters for training L_2-regularized logistic regression with data set S.

Let $\hat{q}' = \arg\max_q J'(q)$ be the parameters for training with data set S'=MS.

Then $J(\hat{q}) = J'(M\hat{q})$

$\hat{q} = M^T\hat{q}'$

$\hat{q}' = M\hat{q}$

and

$$L[S](x) = \frac{1}{1+\exp(-\hat{q}^T x)}$$

$$= \frac{1}{1+\exp(-(M\hat{q})^T (Mx))}$$

$$= \frac{1}{1+\exp(-(\hat{q}')^T x')}$$

$$= L[S'](x')$$

25

Other rotationally invariant algorithms

Several other learning algorithms are also rotationally invariant:

- SVMs with linear, polynomial, RBF, and any other kernel K(x, z) which is a function of only x^Tx, x^Tz, and z^Tz.
- Multilayer back-prop neural networks with weights initialized independently from a spherically-symmetric distribution.
- Logistic regression with no regularization.
- The perceptron algorithm.
- Any algorithm using PCA or ICA for dimensionality reduction, assuming that there is no pre-scaling of all input features to the same variance.

26

Rotational invariance & sample complexity

Theorem 2:

Let L be any rotationally invariant learning algorithm,
$0 < e < 1/8$,
$0 < d < 1/100$,
m is the number of training examples, and
n is the number of features.

Then there exists a learning problem D so that:

(i) The labels depend on a single feature: to y = 1 iff $x_1 = t$, and

(ii) To attain e or lower 0/1 test error with probability at least 1 – d,
L requires a training set of size
$m = O(n/e)$

27

Sketch of theorem 2 proof

For any rotationally invariant learning algorithm L,
$0 < e < 1/8$, and
$0 < d < 1/100$:

Consider the concept class of all n-dimensional linear separators,

$$C = \{h_q : h_q(x) = 1\{q^T x \geq b\}, q \neq 0\}$$

where 1{•} is an indicator function.

The VC-dimension of C is n + 1 (Vapnik, 1982).

From a standard probability approximately correct (PAC) lower bound (Anthony & Bartlett, 1999) it follows that:

For L to attain e or lower 0/1 misclassication error with probability at least 1-d, it is necessary that the training set size be at least
$m = O(n/e)$.

28

Theorem 2 discussion

Rotationally-invariant learning requires a number of training examples that is at least linear in the number of input features.

But, a good feature selection algorithm should be able to learn with O(log n) examples (Ng 1998).

So, rotationally-invariant learning algorithms are ineffective for feature selection in high-dimensional input spaces.

29

Rotational invariance and SVMs

SVMs can classify well with high-dimensional input spaces,
but theorem 2 indicates that they have difficulties with many irrelevant features.

This can be explained by considering both the margin and the radius of the data.

Adding more irrelevant features does not change the margin ?.
However, it does change the radius r of the data.

The expected number of errors in SVMs is a function of $r^2/?^2$ (Vapnik, 1998).
Thus, adding irrelevant features does harm SVM performance.

30

Experimental objectives

The author designed a series of toy experiments to test the theoretical results of this paper.

Three different experiments compare the performance of logistic regression with regularization based on L_1 and L_2 norms.

Each experiment tests different data dimensionalities with a small number of relevant features.

31

Experimental setup

Training and test data are created with a generative logistic model

$$p(y=1|x;q)=1/1+\exp(-q^T x)$$

In each case, inputs x are drawn from a multivariate normal distribution.

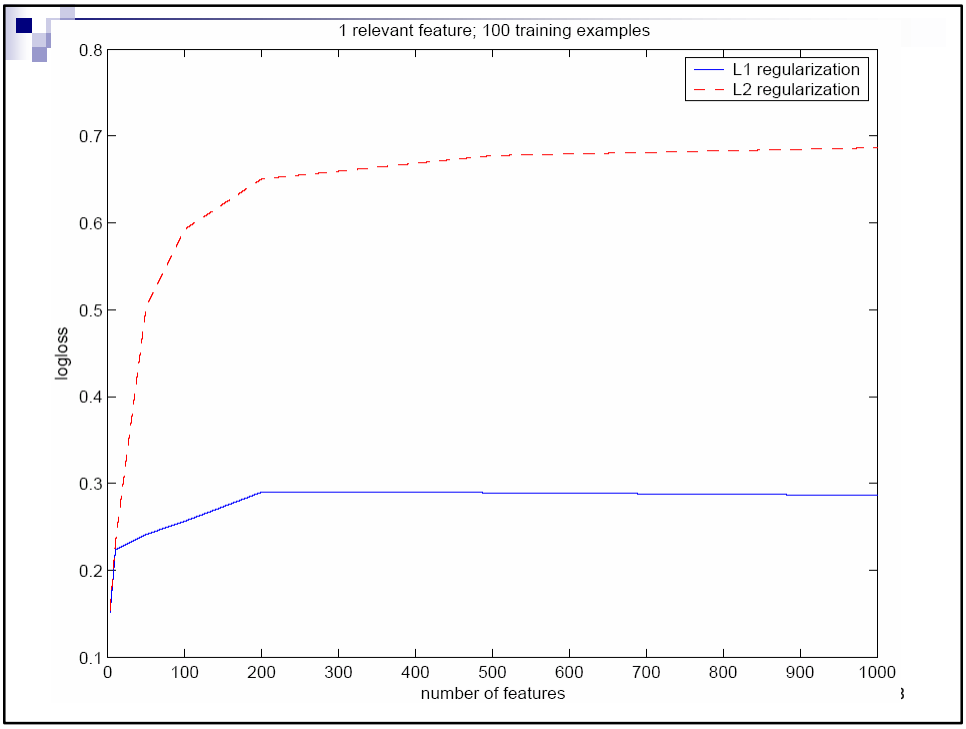
30% of each data set is used as the hold-out set to determine the regularization parameter B.

There are there different experiments:

1. Data has one relevant feature:
   $?_1$ = 10, and all other $?_i$ = 0.
2. Data has three relevant features:
   $?_1 = ?_2 = ?_3$ = 10/v3, and all other $?_i$ = 0.
3. Data is generated with exponentially decaying features:
   $?_i$ = v75 $(1/2)^{i-1}$, (i = 1)

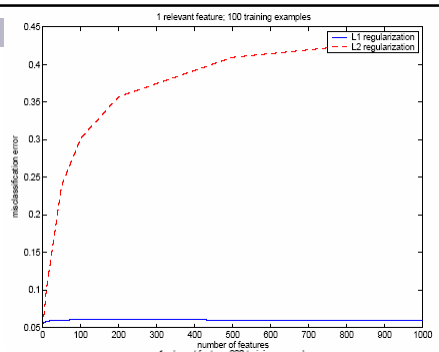
All results are averaged over 100 trials.

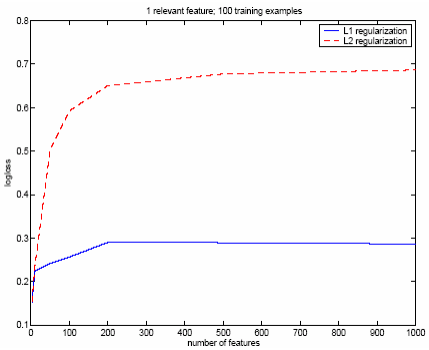
32

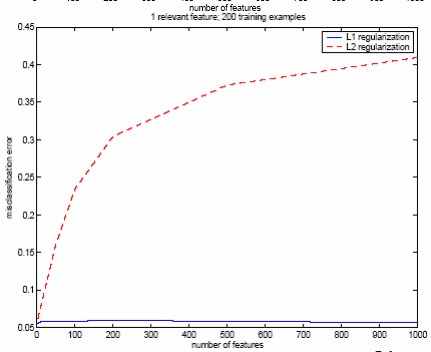


Results

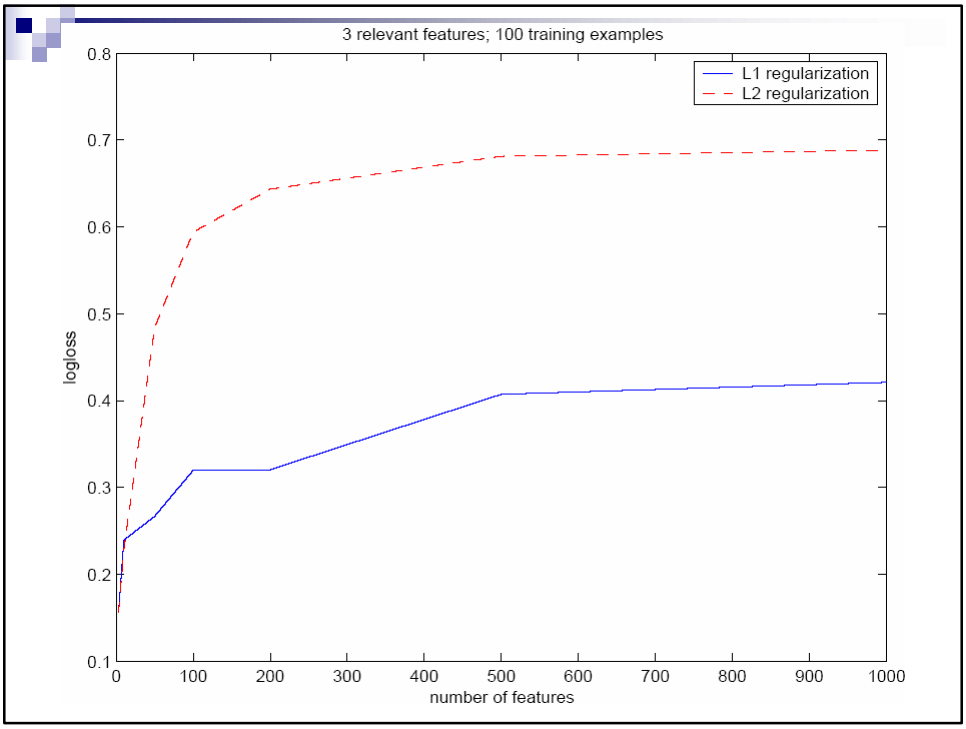
one relevant feature





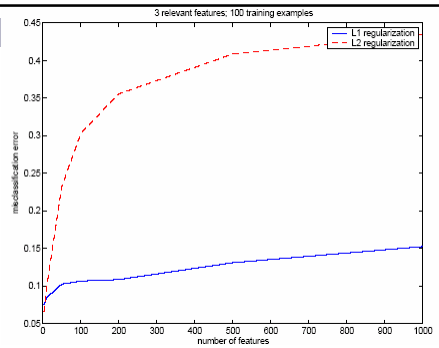


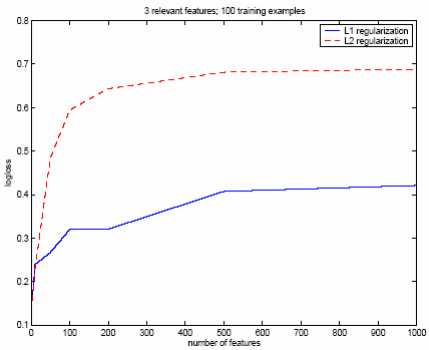
34

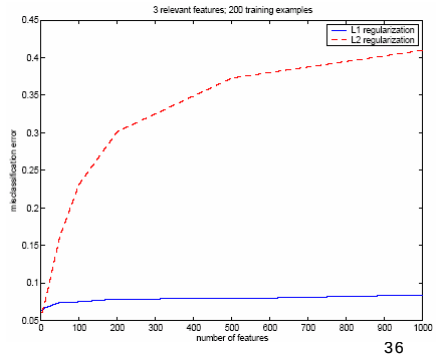


Results

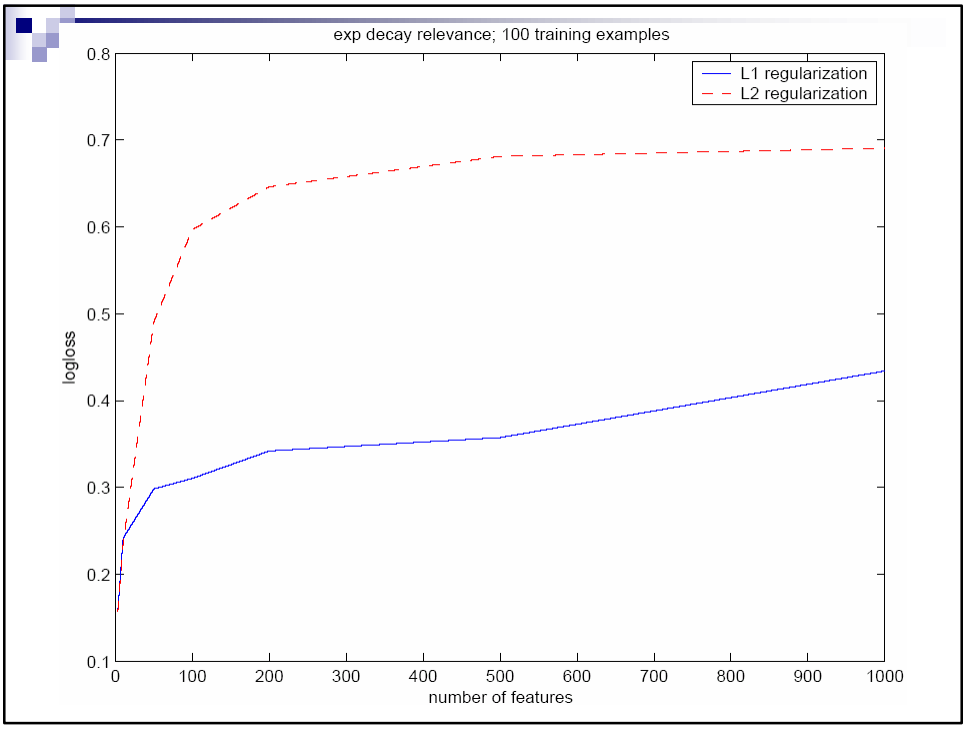
three relevant features





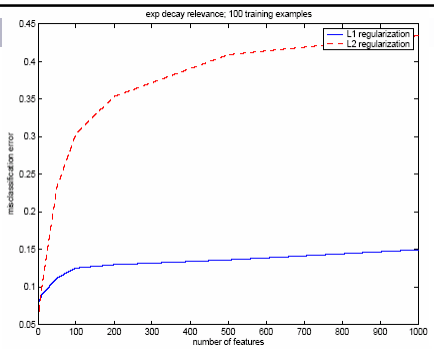


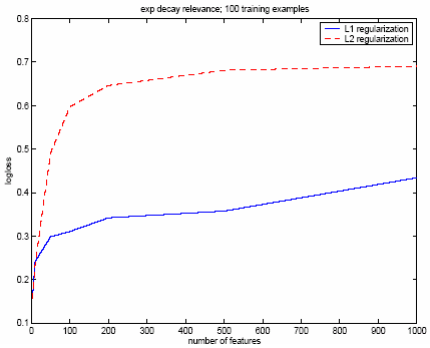
36

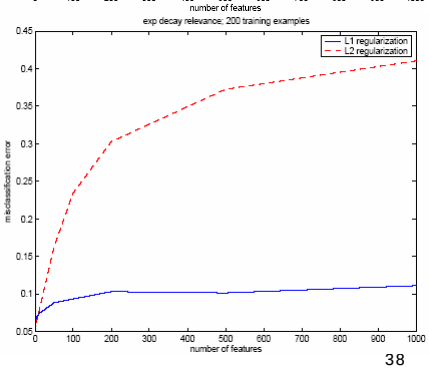


Results

exponential decay of relevance:
$?_i = v75\ (1/2)^{i-1}$, (i = 1)







38

Summary

The paper proves L_1 outperforms L_2 regularization for logistic regression when there are more irrelevant dimensions than training examples.

Experiments show that L_2 regularization classifies poorly for even a few irrelevant features.

Poor performance of L_2 regularization is linked to rotational invariance. Rotational invariance is shared by a large class of other learning algorithms.

These other algorithms presumably have similarly bad performance with many irrelevant dimensions.

References

Anthony, M., & Bartlett, P. (1999). Neural network learning: Theoretical foundations. Cambridge University Press.

Duda, R., Hart, P., Stork, P. (2000). *Pattern Classification, 2nd Ed*. John Wiley & Sons.

Hastie, T., Tibshirani, R., Friedman J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

Jaakkola, T. (2004). Machine Learning lecture notes. Available online at http://people.csail.mit.edu/people/tommi/courses.html

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (second edition)*. Chapman and Hall.

Ng, A. Y. (1998). On feature selection: Learning with exponentially many irrelevant features as training examples. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 404-412). Morgan Kaufmann.

Ng, A. Y. (1998). Machine Learning lecture notes. Available online at http://www.stanford.edu/class/cs229/.

Vapnik, V. (1982). *Estimation of dependences based on empirical data*. Springer-Verlag.

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M. (2003). Use of the Zero-Norm with Linear Models and Kernel Methods. *Journal of Machine Learning Research*, 1439-1461.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 527-550.