

**AN ADAPTIVE REGULARIZATION
CRITERION FOR SUPERVISED
LEARNING**

Dale Schuurmans
Finnegan Southey

Presented by:

Bianca Zadrozny
Department of Computer Science and Engineering
University of California, San Diego

April 11, 2001

Outline

1. The overfitting vs. underfitting dilemma
2. Existing techniques for choosing model complexity
3. Geometry of supervised learning
4. An adaptive regularization criterion
5. Examples: polynomial and radial basis function regression, conditional probability estimation
6. Open issue: 0/1 classification

The overfitting vs. underfitting dilemma

In supervised learning, given training examples $\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle$, we attempt to infer a hypothesis function $h : X \rightarrow Y$ that achieves a small error on future test examples.

Overfitting:

- The hypothesis h is chosen from a class that is too **complex** for the data.
- Test error is large even though training error is small.

Underfitting:

- The hypothesis h is chosen from a class that is too **simple** to capture important structure in the data.
- Training and test errors are large.

We strive to adjust hypothesis complexity given the data so that both underfitting and overfitting are avoided.

Note: complexity is the size of the space of alternative hypotheses considered.

Existing techniques for adjusting hypothesis complexity

- Model selection: a hypothesis class H is decomposed into a discrete collection of subclasses
 $H_0 \subset H_1 \subset \dots \subset H$

The optimal subclass is identified using:

- complexity penalty + training error or statistical selection criteria
- hold-out testing (e.g. cross-validation and bootstrapping) for estimating test error

- Regularization: a penalty is imposed on the individual hypotheses in the base hypothesis class. Examples:

- penalizing the parametric form of the hypothesis (e.g. ridge regression or weight decay)

$$\sum_{i=1}^l (h(x_i) - y_i)^2 + \lambda \sum_{i=1}^l w_i^2$$

- penalizing the global smoothness properties of the hypothesis (e.g. minimizing curvature)

- Model averaging: a composite prediction function is created by taking a weighted combination of many base hypotheses (e.g. bagging and boosting).

One difficulty with these methods is that they generally involve free parameters that are hard to set correctly, e.g., λ in ridge regression.

An adaptive regularization criterion

A hypothesis which fits the training data well but behaves erratically off the training examples is not likely to generalize well.

Instead of minimizing training error alone, we seek hypotheses that behave similarly both on and off the training data.

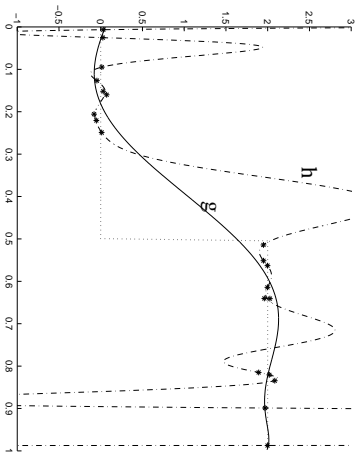


Figure 1: Overfitting effects of polynomial curve fitting

Specifically, we measure the distance the hypothesis exhibits to a fixed function ϕ on both the labeled training data and on the unlabeled training data.

If these distances disagree this indicates that the hypothesis is behaving erratically off the labeled training set.

Geometry of supervised learning

In learning a hypothesis function $h : X \rightarrow Y$ we are interested in modeling the conditional distribution $P_{Y|X}$ given training examples from $P_{X \times Y}$.

Having information about P_X of x from unlabeled data can be useful for choosing a hypothesis.

For any two hypothesis functions f and g we can obtain a measure of the distance between them by computing the expected disagreement in their predictions:

$$d(f, g) \equiv \int \text{err}(f(x), g(x)) dP_X$$

When g is the target conditional distribution, the distance above is the prediction error:

$$d(f, P_{Y|X}) \equiv \int \int \text{err}(f(x), y) dP_{Y|x} dP_X$$

The goal is to find the hypothesis $h \in H$ that is closest to $P_{Y|X}$ using only estimates of $d(h, P_{Y|X})$ given by

$$\hat{d}(h, P_{Y|X}) \equiv \frac{1}{t} \sum_{i=1}^t \text{err}(h(x_i), y_i)$$

where $\langle x_1, y_1 \rangle, \dots, \langle x_t, y_t \rangle$ are training examples.

Training objectives

Two concrete training objectives are proposed:

- Empirical training error plus an additive penalty:

$$\hat{d}(h, P_{Y|X}) + |d(h, \phi) - \hat{d}(h, \phi)|$$

- Empirical training error times a multiplicative penalty:

$$\hat{d}(h, P_{Y|X}) \times \frac{\hat{d}(h, \phi)}{\tilde{d}(h, \phi)}$$

or more precisely

$$\hat{d}(h, P_{Y|X}) \times \max \left(\frac{\hat{d}(h, \phi)}{\tilde{d}(h, \phi)}, \frac{\hat{d}(h, \phi)}{d(h, \phi)} \right)$$

where \hat{d} is computed using labeled examples and d is computed using unlabeled examples.

The multiplicative objective more harshly penalizes discrepancies between on and off training set behavior. Experimentally, it gives the best results.

If the origin function ϕ happens to be $P_{Y|X}$ then minimizing these objectives becomes equivalent to minimizing the true prediction error $d(h, P_{Y|X})$.

This regularization criterion depends on the specific labeled training set under consideration.

7

Regression

In regression problems, $Y = \mathbf{R}$ and we measure prediction error by squared loss $err^m(\hat{y}, y) = (\hat{y} - y)^2$.

We choose a hypothesis to minimize

$$\sum_{i=1}^t (h(x_i) - y_i)^2 / t \times \max \left(\frac{\sum_{j=1}^t (h(x_j) - \phi(x_j))^2 / r}{\sum_{j=1}^t (h(x_j) - \phi(x_j))^2 / r}, \frac{\sum_{i=1}^t (h(x_i) - \phi(x_i))^2 / t}{\sum_{j=1}^t (h(x_j) - \phi(x_j))^2 / r} \right)$$

where $\{x_i, y_i\}$ for $i = 1$ to $i = t$ is the set of labeled training data, $\{x_j\}$ for $j = 1$ to $j = t$ is a large set of unlabeled test examples, and ϕ is a fixed constant (usually zero or constant at the mean of the y labels).

8

Polynomial regression

In polynomial regression, H is the class of polynomials, which can be naturally stratified into the subclasses of polynomials of degree $d = 1, 2, \dots$

It is well-known that fitting data with polynomials can lead to dramatic overfitting.

Polynomial regression: experimental comparison

New ADA method is compared against:

- Model selection: take the best fit polynomials of degree $0, 1, 2, \dots$ and attempt to select the best one.
 - 10-fold cross validation (10CV)
 - structural risk minimization (SRM)
 - complexity penalization and other statistical methods (GCV, AIC, BIC, FPE, CP and RIC)
 - author's previous metric-based model selection strategy (ADJ)
- Regularization methods: consider maximum degree polynomials but penalize individual polynomials based on the size of their coefficients or their smoothness properties.
 - Ridge penalization places a penalty on polynomial coefficients of $\lambda \sum_{k=1}^d a_k^2$ (REG)
 - Bayesian posterior probability maximization with zero mean Gaussian priors on polynomial coefficients a_k with diagonal covariance matrix λY (MAP).

Polynomial regression: experimental results

	target = poly			target = $\sin^2 x$		
	avg	med	std	avg	med	std
ADA Reg	0.077	0.060	0.090	0.107	0.081	0.066
REG opt	0.147	0.082	0.121	0.140	0.092	0.099
MAP opt	0.460	0.099	0.511	0.496	0.232	0.983
ADJ	0.116	0.062	0.188	0.188	0.114	0.150
10CV	0.321	0.065	3.160	0.559	0.132	1.980
SRM	0.163	0.062	1.230	0.576	0.128	2.430
GCV	2421	0.072	4.2e4	4.8e3	0.227	5.6e4

	target = step			target = $\sin x$		
	avg	med	std	avg	med	std
ADA Reg	0.391	0.366	0.113	0.444	0.425	0.085
REG opt	0.371	0.355	0.071	0.429	0.424	0.041
MAP opt	0.496	0.400	0.385	0.651	0.476	0.989
ADJ	0.458	0.466	0.112	0.712	0.504	0.752
10CV	14.90	0.420	340.0	2.410	0.516	14.20
SRM	29.00	0.510	311.0	29.40	0.781	469.0
GCV	3.2e5	51.9	3.1e6	1.4e5	11.3	2.6e6

- ADA outperforms fixed regularization strategies for all fixed λ .
- Since it outperforms even the best choice of λ , ADA shows adaptation to the specific training set.
- The only model selection strategy to perform consistently is ADJ, which also uses unlabeled training data.
- Even cross validation did not perform well.

11

Radial basis function regression

Given a set of prototype centers c_1, \dots, c_k a radial basis function (RBF) representation of a hypothesis h is given by

$$h(x) = \sum_{i=1}^k w_i g\left(\frac{\|x - c_i\|}{\sigma}\right)$$

where $\|x - c_i\|$ is the Euclidean distance between x and center c_i and g is a response function with width parameter σ . Here we use $g(z) = e^{-z^2}$.

The simplest approach to function fitting is to place a prototype center on each training example and determine the weight vector w by solving the equation above for w .

Fitting functions with RBF networks also has the problem that the training data is generally overfit.

12

Radial basis function: experimental comparison

We compare the adaptive regularization criterion to adding a ridge penalty to the weight vector, therefore minimizing

$$\sum_{i=1}^l (h(x_i - y_i))^2 + \lambda \sum_{i=1}^l w_i^2$$

To apply this method in practice one has to make an intelligent choice of λ and σ .

Radial basis function: experimental results

Each dataset is randomly split into training, unlabeled, and test sets.

The random dataset split is repeated 100 times.

Data set	ADA	REG opt
AAUP	0.0197 ± 0.004	0.0361 ± 0.009
BOSTON-C	0.150 ± 0.0212	0.155 ± 0.0197
BODYFAT	0.131 ± 0.0171	0.129 ± 0.0150
ABALONE	0.034 ± 0.0046	0.050 ± 0.0055

- ADA performs better than any fixed regularizer on every problem.
- ADA beats REG-opt in all but one problem, showing that ADA chooses the regularization parameters adaptively based on the given training set.

Conditional density estimation

Consider a setting where hypotheses make probabilistic predictions, $h(x) \in [0, 1]$ of y labels in $0, 1$.

In this situation, we measure distances between conditional probability models using the KL-divergence:

$$d(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} - (1 - f(x)) \log \frac{1 - f(x)}{1 - g(x)} dP_X$$

The goal is to minimize the KL-divergence $d(P_{Y|X}\|h)$, which amounts to minimizing the log-loss:

$$err(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Even trivial probability models can overfit sensible training data, that is, we can get a small log-loss on the training data $\hat{d}(P_{Y|X}\|h)$ and yet get a large log-loss on test examples.

Conditional density estimation: an example

Consider a one-dimensional logistic regression

$$h(x) = \frac{1}{1 + e^{-ax-b}}$$

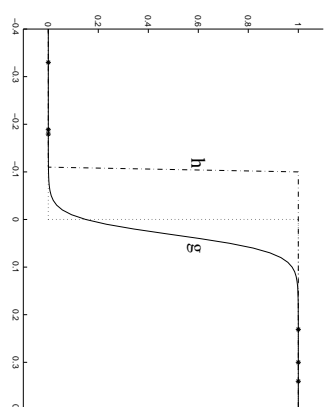
for parameters a and b . This defines a “smoothed” step function with a crossover value at $x = -\frac{b}{a}$.

The naive approach is choosing a and b to minimize the log-loss on the training data, but this is likely to overfit.

We can regularize the hypothesis h by minimizing

$$\hat{d}(P_{Y|X}\|h) \times \max \left(\frac{d(P_{Y|X}\|h)}{\hat{d}(P_{Y|X}\|h)}, \frac{\hat{d}(P_{Y|X}\|h)}{d(P_{Y|X}\|h)} \right)$$

using the origin function $\phi(x) = \frac{1}{2}$.



$$\begin{aligned} \hat{d}(P_{Y|X}\|h) &= 0 \\ d(P_{Y|X}\|h) &= 30.6 \\ \hat{d}(P_{Y|X}\|g) &= 4 \times 10^{-6} \\ d(P_{Y|X}\|g) &= 0.0249 \end{aligned}$$

Figure 2: Maximum likelihood (h) versus regularized (g) logistic regression.

Open issue: classification

The adaptive regularization scheme can potentially be used in classification.

In this case, distances are measured by the disagreement probability:

$$d(f, g) = \int \mathbf{1}_{(f(x) \neq g(x))} dP_X = P_X(f(x) \neq g(x))$$

However, preliminary results on decision tree pruning are inconclusive, and it appears that the technique does not work as decisively for classification problems.

The difficulty is that classification functions are histogram-like and behave similarly in large neighborhoods around training points, so distances on labeled and unlabeled data points are similar even for complex models.