April 11, 2001

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

Presented by:

Prineeta Southey
Dale Schuurmans

LEARNING CRITERION FOR SUPERVISED
AN ADAPTIVE REGULARIZATION
\( \gamma \) in ridge regression.

One difficulty with these methods is that they generally

\[ \min \{ \sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \} \]

\( \alpha \) is the size of the space of alternatives.

We strive to adjust hypotheses complexity given the data.

\[ \text{Training and test errors are large.} \]

\[ \text{The hypotheses in the base hypotheses class.} \]

\[ \text{Uneventful:} \]

\[ \text{Test error is large even though training error is small.} \]

\[ \text{complex for the data.} \]

\[ \text{The hypotheses } f \text{ is chosen from a class that is too} \]

\[ \text{overfitting.} \]

\[ \text{Select a small error on the test examples.} \]

In supervised learning, \( f \) is trained examples (\( x \), \( y \)) in the space of hypotheses \( \mathcal{H} \).
where \( \langle \theta, x \rangle \) are training examples.

\[
\langle \theta, x \rangle \sim \frac{1}{Z} e^{\langle x, \phi \rangle} \prod_i \left( 1 + \frac{1}{\alpha} \langle \phi, f \rangle \right)
\]

to find the hypotheses \( \theta \) that is closer

\[
\mathcal{L}_{DP} \equiv \frac{1}{\mathcal{L}_{DP}} \mathcal{L}_{DP} \mathcal{L}_{DP}(\theta, \phi, f) \mathcal{L}_{DP} \mathcal{L}_{DP} \mathcal{L}_{DP}(\theta, \phi, f) \mathcal{L}_{DP}
\]

above is the prediction error.

When \( \theta \) is the larger conditional distribution, the distance

\[
\mathcal{L}_{DP} \equiv \frac{1}{\mathcal{L}_{DP}} \mathcal{L}_{DP} \mathcal{L}_{DP}(\theta, \phi, f) \mathcal{L}_{DP} \mathcal{L}_{DP} \mathcal{L}_{DP}(\theta, \phi, f) \mathcal{L}_{DP}
\]

expected disagreement between them by combining the
disagreement between them can obtain a

for any two hypotheses functions \( f \) and \( g \) we can obtain a

can be used for choosing a hypotheses

Having information about \( \mathcal{L}_{DP} \) of \( x \) from unlabeled data

1. \( \mathcal{L}_{DP} \) from labeled examples from \( \mathcal{L}_{DP} \)
2. \( \mathcal{L}_{DP} \) in modeling the conditional distribution \( \mathcal{L}_{DP} \)

In learning a hypothesis function \( \eta \) we are

Geometry of Supervised Learning

An adaptive regularization criterion
We choose a hypothesis to minimize

\[ \mathcal{L}(\hat{f} - f) = (\hat{f}, f)^T \text{corr}(\hat{f}, f) \]

In regression problems, \( \text{corr} \) and \( \text{mse} \) are proposed.

**Regression**

We predict error from unlabeled examples.

The regularization criterion depends on the specific label.

The prediction error \( p, q \).

If the original function happens to be \( \lambda^2 \), then minimizing the multiplier, it gives the best results. The multiplior is motivated by an off training set behavior.

The multiplior objective more handly penalizes 3.

Where \( p \) is computed using labeled examples and \( q \) is computed using unlabeled examples.
MAP with diagonal covariance matrix $\Sigma$'s variance = 0

Bayesian posterior probability maximization

\[
\hat{\beta} = \arg\max_{\beta} \text{prob}(\beta|X,Y)
\]

Kriging approximation

Properties of polynomials:

- Regularization methods: consider maximum degree

\[e^y(A\beta)\]

- authors' previous meta-based model selection strategy:
  - odds (CVA, RIC, CP, and RIC)
  - cross-validation and other statistical methods

- Internal risk minimization (SM)
- 10-fold cross validation (10CV)

- $p = 0.1$, $\ldots$, $P$ and attempt to select the best one.

- Model selection: take the best fit of polynomials of de-

- New ADA method is compared against:

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

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>CV</th>
<th>3.25</th>
<th>3.45</th>
<th>1.45</th>
<th>1.35</th>
<th>2.65</th>
<th>2.36</th>
<th>3.19</th>
<th>3.36</th>
<th>2.76</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
<tr>
<td>C</td>
<td>CV</td>
<td>3.25</td>
<td>3.45</td>
<td>1.45</td>
<td>1.35</td>
<td>2.65</td>
<td>2.36</td>
<td>3.19</td>
<td>3.36</td>
<td>2.76</td>
<td>3.06</td>
</tr>
</tbody>
</table>

Poly neural network expression & parameters
- x

Radial basis function regression
- Given a set of prototype centers, a radial basis
  function (RBF) representation of a hypothesis is given.
- The weight vector $\omega$ by solving the equation above for $m$. The simplest approach to function fitting is to place a
  where $||x - \omega||^2 = (x)\eta$
The random dataset split is repeated 100 times.

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

Radial basis function experimental results

To apply this method in practice one has to make an in-

$\frac{1}{m} \sum_{i=1}^{m} y + \frac{1}{2} (f - x)^2 \psi(x)$

A ridge penalty to the weight vector, therefore minimizing

We compared the adaptive regularization criterion to adding

Radial basis function experimental comparison
Figure 2: Maximum likelihood (ML) versus maximum (MLE) estimation.

\[
\hat{\nu} = (x) \phi
\]

using the optimal function

\[
\frac{\left(\phi \left| \nu \text{KL} \right| d \nu \right)}{(\phi \left| \nu \text{KL} \right| d \nu)} \max \times \left(\phi \left| \nu \text{KL} \right| d \nu \right)
\]

We can regularize the hypothesis by minimizing log-likelihood on the training data, but this is tricky to optimize. The naive approach is choosing \( a \) and \( q \) to minimize the function with a crossover value at \( x \).

for parameters \( a \) and \( q \). This defines a "smoothed" step

\[
\frac{q - x - 2}{1} = (x) \phi
\]

Consider a one-dimensional logistic regression

example.

If you get a large log-loss on test data, then is we can get a small log-loss on the training data.

Even trivial probability models can overfit sensible training data.

\[
(\nu - 1) \log (\nu - 1) - \nu \log \nu - (\nu - 1) \eta \log (\nu - 1) \eta = (\nu - 1) \eta \log (\nu - 1) \eta
\]

where \( \text{ momentum } \) is minimized the log-loss:

The goal is to minimize the KL-divergence.

The probability models using the KL-divergence:

In this situation, we measure distance between conditional predictions, \( \phi \) (\( y \) on \( x \) in 0, 1, conditional density estimation an example)
points are similar even for complex models. Training points, as distances on labeled and unlabeled data, accurately predict. In some cases, modeling distances around the few labels does not imply that the technique does not work as described for classification problems. However, preliminary results on decision tree pruning are

\[(x| \theta \neq (x)f)_d = \gamma D p((x| \theta \neq (x)f) f = (\theta| f)p\]

In this case, distances are measured by the disagreement in classification.

The adaptive regularization scheme can potentially be used for Open issue classification.