Differentiable Sparse Coding

David Bradley and J. Andrew Bagnell
NIPS 2008
100,000 ft View

- Complex Systems

Joint Optimization

Cameras

Ladar

Voxel Features

Voxel Classifier

Voxel Grid

Column Cost

2-D Planner

Y (Path to goal)

Initialize with “cheap” data

Joint Optimization
10,000 ft view

- Sparse Coding = Generative Model

\[ \hat{X} \rightarrow \text{Optimization} \rightarrow \hat{W} \rightarrow \text{Classifier} \]

Unlabeled Data

- Semi-supervised learning
- KL-Divergence Regularization
- Implicit Differentiation
Sparse Coding

Understand X
As a combination of factors
Sparse coding uses optimization

$$\tilde{x} \approx f (B\tilde{w})$$

Projection (feed-forward):

$$\tilde{w} = f (\tilde{x}^T B)$$

Reconstruction loss function

Some vector

Want to use to classify $x$
Sparse vectors:
Only a few significant elements
Example: $X = \text{Handwritten Digits}$

\[ \sum \]

Basis vectors colored by $w$

\[ f(B\vec{w}) \approx X \]

\[ 2 \sim 2 \]
Optimization vs. Projection

Input

Projection

Basis
Optimization vs. Projection

Outputs are sparse for each example
Generative Model

\[ X_i = \begin{array}{c} 2 \end{array} \]

\[
P(X) = \int_B \int_W P(X|W,B)P(W)P(B)\,dW\,dB
\]

\[
P(X) = \int_B P(B) \int_W \prod_i P(X_i|W_i,B) P(W_i)\,dW\,dB
\]

- Latent variables are Independent
- Likelihood
- Prior
- Examples are Independent

11
Sparse Approximation

\[ X = 2 \]

\[ P(\vec{x}) \approx \max_{\vec{w}} P(\vec{x} | \vec{w}, B) P(\vec{w}) \]

\[ \hat{\vec{w}} = \arg \max_{\vec{w}} P(\vec{x} | \vec{w}, B) P(\vec{w}) \]

MAP Estimate
Sparse Approximation

\[ \arg \max_{\tilde{w}} P(\tilde{x}) = \arg \min_{\tilde{w}} \left( - \log P(\tilde{x}) \right) \]

\[ - \log P(\tilde{x}) \propto \text{Loss}(\tilde{x} \| f(B\tilde{w})) + \lambda \text{Prior}(\tilde{w} \| \bar{p}) \]

Distance between reconstruction and input

Distance between weight vector and prior mean

Regularization Constant
Example: Squared Loss + L1

\[ \hat{w} = \arg \min_{\vec{w}} \sum_i (x_i - B^i \vec{w})^2 + \lambda \sum_i |\vec{w}_i| \]

• Convex + sparse (widely studied in engineering)
• Sparse coding solves for B as well (non-convex for now...)
• Shown to generate useful features on diverse problems

Donoho and Elad, *Proceedings of the National Academy of Sciences*, 2002
Raina, Battle, Lee, Packer, Ng, *ICML*, 2007
L1 Sparse Coding

\[ \arg \min_{W,B} \sum_{j} \frac{1}{2} \| Bw_j - x \|_2^2 + \lambda \| w_j \|_1 \]

Optimize B over all examples

\[ s.t. \quad \| b_j \|_2 = 1 \]

Shown to generate useful features on diverse problems
Differentiable Sparse Coding


Raina, Battle, Lee, Packer, Ng, ICML, 2007
L1 Regularization is Not Differentiable

Why is this unsatisfying?
Problem #1: Instability

- L1 Map Estimates are discontinuous
- Outputs are not differentiable
- Instead use KL-divergence

Proven to compete with L1 in online learning
Problem #2: No closed-form Equation

\[ \hat{w} = \operatorname{arg\ min}_{\hat{w}} \operatorname{Loss}(\hat{x} \parallel f(B\hat{w})) + \lambda \operatorname{Prior}(\hat{w} \parallel \tilde{p}) \]

At the MAP estimate:

\[ \nabla_{\hat{w}} \operatorname{Loss}(\hat{x} \parallel f(B\hat{w})) + \lambda \nabla_{\hat{w}} \operatorname{Prior}(\hat{w} \parallel \tilde{p}) = 0 \]
Solution: Implicit Differentiation

Differentiate both sides with respect to an element of B:

\[
\frac{\partial}{\partial B^i_j} \left( \nabla_w \text{Loss}(\tilde{x} \mid f(B\hat{w})) + \lambda \nabla_w \text{Prior}(\hat{w} \mid \bar{p}) \right) = 0
\]

Since \( \hat{w} \) is a function of B:

\[
\frac{\partial}{\partial B^i_j} (\hat{w}) = \frac{\partial \hat{w}}{\partial B^i_j}
\]
Example: Squared Loss, KL prior

$$\hat{w} = \arg\min_w \frac{1}{2} \|x - Bw\|_2^2 + \lambda \sum_i w_i \log \frac{w_i}{p_i} - w_i + p_i$$

KL-Divergence

$$\frac{\partial \hat{w}}{\partial B^k_i} = - \left( B^T B + \text{diag} \left( \frac{\lambda}{\hat{w}} \right) \right)^{-1} \left( (B^k \hat{w}_i)^T + \bar{e}_i (f(B^k \hat{w}) - x_k) \right)$$
Handwritten Digit Recognition

50,000 digit training set
10,000 digit validation set
10,000 digit test set
Handwritten Digit Recognition

Step #1:

Unsupervised Sparse Coding
$L_2$ loss and $L_1$ prior

Training Set

Raina, Battle, Lee, Packer, Ng, ICML, 2007
Handwritten Digit Recognition

Step #2:

Maxent Classifier

Loss Function

Sparse Approximation

0123456789
Handwritten Digit Recognition

Step #3:

Maxent Classifier

Loss Function

Supervised Sparse Coding

0123456789
KL Maintains Sparsity

![Graph showing sparsity maintenance](image)

- Log scale
- Weight concentrated in few elements

Activation Magnitude vs Basis Vector (sorted by activation)
KL adds Stability

Performance vs. Prior

![Bar chart showing misclassification error (%) vs. number of training examples for L1, KL, and KL + Backprop methods.](chart.png)

Better
Classifier Comparison

![Bar chart comparing misclassification error for different classifiers and methods: Maxent, 2-layer NN, SVM (Linear), SVM (RBF) with PCA, L1, KL, and KL+backprop. The chart shows that SVM (RBF) with KL+backprop has the lowest misclassification error among the options.](image-url)
Comparison to other algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>L1</th>
<th>KL</th>
<th>KL+backprop</th>
<th>SVM</th>
<th>2-layer NN [15]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set Error</td>
<td>3.53%</td>
<td>2.21%</td>
<td>1.30%</td>
<td>1.4%</td>
<td>1.6%</td>
</tr>
</tbody>
</table>
Transfer to English Characters

24,000 character training set
12,000 character validation set
12,000 character test set

8

16

[Image: Unexpected]
Transfer to English Characters

Step #1:

[Diagram showing the process from Digits Basis to Sparse Approximation, then to Maxent Classifier and finally to Loss Function.]

Raina, Battle, Lee, Packer, Ng, ICML, 2007
Transfer to English Characters

Step #2:

Maxent Classifier

Loss Function

Supervised Sparse Coding

Unexpected
Transfer to English Characters

Classification Accuracy

Better

Training Set Size

500 5000 20000

- Raw
- PCA
- L1
- KL
- KL+backprop
Text Application

Step #1:
5,000 movie reviews

10 point sentiment scale
1=hated it, 10=loved it

Unsupervised Sparse Coding
KL loss + KL prior

Pang, Lee, Proceeding of the ACL, 2005
Text Application

Step #2:
5-fold Cross Validation

10 point sentiment scale
1=hated it, 10=loved it
Text Application

Step #3:
5-fold Cross Validation

10 point sentiment scale
1=hated it, 10=loved it
Movie Review Sentiment

State of the art graphical model

Supervised basis

Unsupervised basis

Predictive $R^2$

Better

Blei, McAuliffe, NIPS, 2007

- LDA
- KL
- sLDA
- KL+backprop
Future Work

Camera

Engineered Features

Sparse Coding

MMP

Voxel Classifier

Example Paths

Labeled Training Data

Laser

RGB Camera

NIR Camera

Ladar

Example Paths

Labeled Training Data
Future Work: Convex Sparse Coding

• Sparse approximation is convex
• Sparse coding is not because fixed-size basis is a non-convex constraint
• Sparse coding $\leftrightarrow$ sparse approximation on infinitely large basis + non-convex rank constraint
  – Relax to a convex $L_1$ rank constraint
• Use boosting for sparse approximation directly on infinitely large basis

Bengio, Le Roux, Vincent, Dellalleau, Marcotte, NIPS, 2005
Zhao, Yu. Feature Selection for Data Mining, 2005
Questions?