Making Do with Less: An Introduction to Compressed Sensing 5
Classification

Kurt Bryan

July 12, 2023

Scenario

A note (bill) is inserted into a vending machine, illuminated with various LED's, spectral data collected.

Scenario

A note (bill) is inserted into a vending machine, illuminated with various LED's, spectral data collected.

The data collected can be arranged in a vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_N).$$

Scenario

A note (bill) is inserted into a vending machine, illuminated with various LED's, spectral data collected.

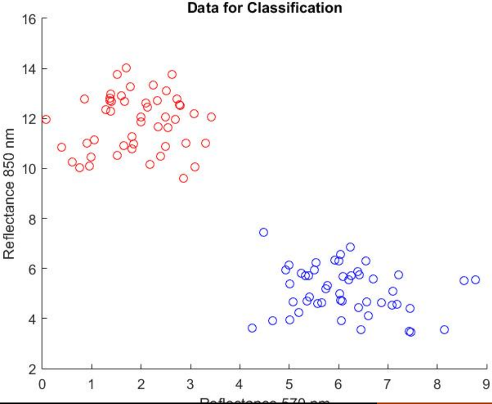
The data collected can be arranged in a vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_N).$$

Goal: Develop an algorithm that uses $\mathbf{x}$ (the "feature vector") to determine the national origin and denomination of the note—*classify* the note.

Example

Two classes: US \$1 and \$20 notes, $x_1 =$ reflectance at 570 nm, $x_2 =$ reflectance at 850 nm. Fifty copies of each note are inserted into a bill validator and reflectance data collected:



Training

A common framework for developing a classifier involves *supervised learning*. If there are only two classes

Training

A common framework for developing a classifier involves *supervised learning*. If there are only two classes

1. Collect *training data*:
   - Feature vectors $\mathbf{x}_1^k$, $1 \leq k \leq C_1$ for notes in class 1
   - Feature vectors $\mathbf{x}_2^k$, $1 \leq k \leq C_2$ for notes in class 2.

Training

A common framework for developing a classifier involves *supervised learning*. If there are only two classes

1. Collect *training data*:
   - Feature vectors $\mathbf{x}_1^k$, $1 \leq k \leq C_1$ for notes in class 1
   - Feature vectors $\mathbf{x}_2^k$, $1 \leq k \leq C_2$ for notes in class 2.
2. Find a function $\phi : \mathbb{R}^N \to \mathbb{R}$ such that
   - $\phi(\mathbf{x}_1^k) < 0$ for notes in class 1
   - $\phi(\mathbf{x}_2^k) > 0$ for notes in class 2.

   This is the *training*.

Training

A common framework for developing a classifier involves *supervised learning*. If there are only two classes

1. Collect *training data*:
   - Feature vectors $\mathbf{x}_1^k$, $1 \leq k \leq C_1$ for notes in class 1
   - Feature vectors $\mathbf{x}_2^k$, $1 \leq k \leq C_2$ for notes in class 2.
2. Find a function $\phi : \mathbb{R}^N \rightarrow \mathbb{R}$ such that
   - $\phi(\mathbf{x}_1^k) < 0$ for notes in class 1
   - $\phi(\mathbf{x}_2^k) > 0$ for notes in class 2.

   This is the *training*.

A future note with data $\mathbf{x}^*$ can be classified by computing $\phi(\mathbf{x}^*)$.

Training

A common framework for developing a classifier involves *supervised learning*. If there are only two classes

1. Collect *training data*:
   - Feature vectors $\mathbf{x}_1^k$, $1 \leq k \leq C_1$ for notes in class 1
   - Feature vectors $\mathbf{x}_2^k$, $1 \leq k \leq C_2$ for notes in class 2.
2. Find a function $\phi : \mathbb{R}^N \to \mathbb{R}$ such that
   - $\phi(\mathbf{x}_1^k) < 0$ for notes in class 1
   - $\phi(\mathbf{x}_2^k) > 0$ for notes in class 2.

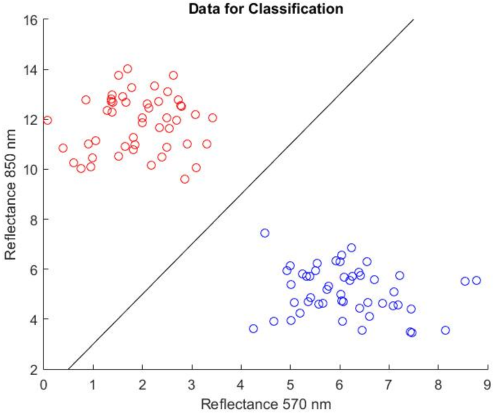
   This is the *training*.

A future note with data $\mathbf{x}^*$ can be classified by computing $\phi(\mathbf{x}^*)$.

The function ϕ might be linear or nonlinear, implemented in a variety of ways.

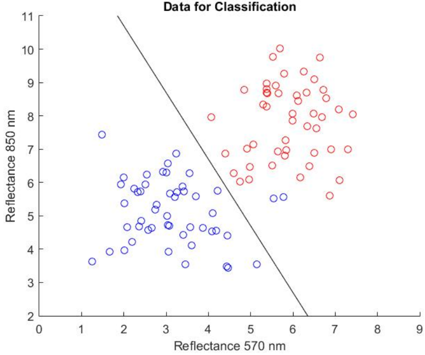
Example

This training data is *linearly separable* by the line $x_2 = 2x_1 + 1$, so we can use $\phi(\mathbf{x}) = x_2 - 2x_1 - 1$.



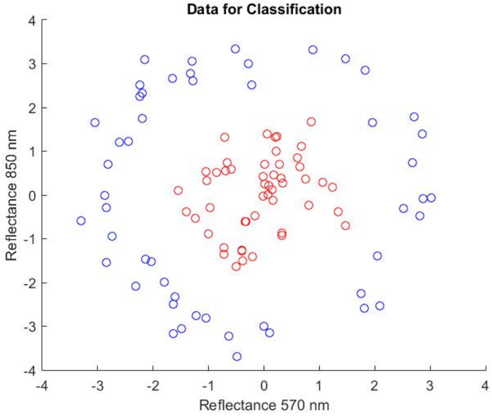
Noise

But the data might not be perfectly linearly separable due to noise, outliers, or the nature of the problem:



Nonlinear Classification

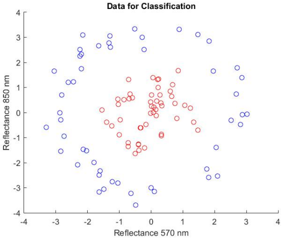
The training data may not be linearly separable at all:



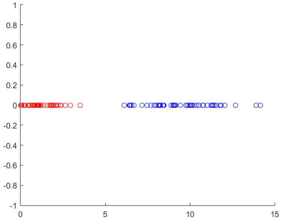
Nonlinear Classification

One could apply a nonlinear transformation $\mathbf{x} \rightarrow \psi(\mathbf{x})$ where $\psi : \mathbb{R}^N \rightarrow V$ to the data, where V is some other space. We can then try a classifier in V, where the data may be linearly separable. V may have dimension other than N.

Example: $\psi(\mathbf{x}) = x_1^2 + x_2^2$.



$\xrightarrow{\psi}$



Classification Techniques

Classification Techniques

- There are many methods for linear classification: support vector machines, linear discriminant analysis, perceptrons.

Classification Techniques

- There are many methods for linear classification: support vector machines, linear discriminant analysis, perceptrons.
- Some of these techniques can be adapted for non-linearly separable data; other methods address nonlinearity directly.

Classification Techniques

- There are many methods for linear classification: support vector machines, linear discriminant analysis, perceptrons.
- Some of these techniques can be adapted for non-linearly separable data; other methods address nonlinearity directly.
- Binary classification techniques can be adapted to handle multiple classes (e.g., a “one versus all” approach.)

A Different Approach: Sparsity and Dictionary Learning

Consider data consisting of feature vectors $\mathbf{x} \in \mathbb{R}^2$ in one of two classes, of the forms

$$\mathbf{x} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ or } \mathbf{x} = k \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

for scalar k (each class lies on a line). Our measurement of $\mathbf{x}$ may contain noise or other errors.

A Different Approach: Sparsity and Dictionary Learning

Consider data consisting of feature vectors $\mathbf{x} \in \mathbb{R}^2$ in one of two classes, of the forms

$$\mathbf{x} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ or } \mathbf{x} = k \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

for scalar k (each class lies on a line). Our measurement of $\mathbf{x}$ may contain noise or other errors.

Suppose we measure feature vector

$$\mathbf{x}^* = \begin{bmatrix} 2.06 \\ -2.12 \end{bmatrix}.$$

To which class does $\mathbf{x}^*$ belong?

Sparsity and Dictionary Learning

Form a "dictionary"

$$\mathbf{D} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with atoms (columns) that are representatives of each class.

Sparsity and Dictionary Learning

Form a "dictionary"

$$\mathbf{D} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with atoms (columns) that are representatives of each class.

Seek a solution to the linear system

$$\mathbf{D} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2.06 \\ -2.12 \end{bmatrix}.$$

Sparsity and Dictionary Learning

Form a "dictionary"

$$\mathbf{D} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with atoms (columns) that are representatives of each class.

Seek a solution to the linear system

$$\mathbf{D}\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2.06 \\ -2.12 \end{bmatrix}.$$

The solution is $\boldsymbol{\alpha}_1 = -0.03, \alpha_2 = 2.09$.

Sparsity and Dictionary Learning

Form a "dictionary"

$$\mathbf{D} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with atoms (columns) that are representatives of each class.

Seek a solution to the linear system

$$\mathbf{D} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2.06 \\ -2.12 \end{bmatrix}.$$

The solution is $\boldsymbol{\alpha}_1 = -0.03, \alpha_2 = 2.09$. We could classify $\mathbf{x}^*$ in class 2 based on $|\alpha_2|$ being much larger than $|\alpha_1|$, but this depends on the scaling of the dictionary columns.

An Alternate Approach: Example 1

Instead of using $|\alpha_2| > |\alpha_1|$ to classify, decompose

$$\boldsymbol{\alpha} = \langle -0.03, 2.09 \rangle = \underbrace{\langle -0.03, 0 \rangle}_{\boldsymbol{\alpha}^1} + \underbrace{\langle 0, 2.09 \rangle}_{\boldsymbol{\alpha}^2}$$

(a "Class 1" and "Class 2" portion).

An Alternate Approach: Example 1

Instead of using $|\alpha_2| > |\alpha_1|$ to classify, decompose

$$\boldsymbol{\alpha} = \langle -0.03, 2.09 \rangle = \underbrace{\langle -0.03, 0 \rangle}_{\boldsymbol{\alpha}^1} + \underbrace{\langle 0, 2.09 \rangle}_{\boldsymbol{\alpha}^2}$$

(a "Class 1" and "Class 2" portion). Compute vectors

$$\mathbf{v}_1 = \mathbf{D}\boldsymbol{\alpha}^1 = \begin{bmatrix} -0.03 \\ -0.03 \end{bmatrix}, \quad \mathbf{v}_2 = \mathbf{D}\boldsymbol{\alpha}^2 = \begin{bmatrix} 2.09 \\ -2.09 \end{bmatrix}.$$

An Alternate Approach: Example 1

Instead of using $|\alpha_2| > |\alpha_1|$ to classify, decompose

$$\boldsymbol{\alpha} = \langle -0.03, 2.09 \rangle = \underbrace{\langle -0.03, 0 \rangle}_{\boldsymbol{\alpha}^1} + \underbrace{\langle 0, 2.09 \rangle}_{\boldsymbol{\alpha}^2}$$

(a "Class 1" and "Class 2" portion). Compute vectors

$$\mathbf{v}_1 = \mathbf{D}\boldsymbol{\alpha}^1 = \begin{bmatrix} -0.03 \\ -0.03 \end{bmatrix}, \quad \mathbf{v}_2 = \mathbf{D}\boldsymbol{\alpha}^2 = \begin{bmatrix} 2.09 \\ -2.09 \end{bmatrix}.$$

Then

$$\|\mathbf{x}^* - \mathbf{v}_1\|_2 = 2.96, \quad \|\mathbf{x}^* - \mathbf{v}_2\|_2 = 0.042.$$

An Alternate Approach: Example 1

Instead of using $|\alpha_2| > |\alpha_1|$ to classify, decompose

$$\boldsymbol{\alpha} = \langle -0.03, 2.09 \rangle = \underbrace{\langle -0.03, 0 \rangle}_{\boldsymbol{\alpha}^1} + \underbrace{\langle 0, 2.09 \rangle}_{\boldsymbol{\alpha}^2}$$

(a "Class 1" and "Class 2" portion). Compute vectors

$$\mathbf{v}_1 = \mathbf{D}\boldsymbol{\alpha}^1 = \begin{bmatrix} -0.03 \\ -0.03 \end{bmatrix}, \quad \mathbf{v}_2 = \mathbf{D}\boldsymbol{\alpha}^2 = \begin{bmatrix} 2.09 \\ -2.09 \end{bmatrix}.$$

Then

$$\|\mathbf{x}^* - \mathbf{v}_1\|_2 = 2.96, \quad \|\mathbf{x}^* - \mathbf{v}_2\|_2 = 0.042.$$

Since $\mathbf{x}^*$ can be built from the class 2 atom much more accurately we assign $\mathbf{x}^*$ to class 2.

An Alternate Approach: Example 2

Consider feature vector in $\mathbb{R}^4$ with three classes.

An Alternate Approach: Example 2

Consider feature vector in $\mathbb{R}^4$ with three classes. Let

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 2 & 1 & 2 & -1 & 1 & 3 & 1 \\ 1 & 0 & -1 & 1 & 1 & -2 & 1 & 0 & 2 \\ 3 & -1 & 0 & 0 & 1 & 1 & 2 & -3 & 1 \\ -5 & 2 & -1 & 3 & 3 & -6 & 0 & -3 & 0 \end{bmatrix}.$$

The first three atoms are class 1, the second three class 2, the last three class 3.

An Alternate Approach: Example 2

Consider feature vector in $\mathbb{R}^4$ with three classes. Let

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 2 & 1 & 2 & -1 & 1 & 3 & 1 \\ 1 & 0 & -1 & 1 & 1 & -2 & 1 & 0 & 2 \\ 3 & -1 & 0 & 0 & 1 & 1 & 2 & -3 & 1 \\ -5 & 2 & -1 & 3 & 3 & -6 & 0 & -3 & 0 \end{bmatrix}.$$

The first three atoms are class 1, the second three class 2, the last three class 3.

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$.

An Alternate Approach: Example 2

Consider feature vector in $\mathbb{R}^4$ with three classes. Let

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 2 & 1 & 2 & -1 & 1 & 3 & 1 \\ 1 & 0 & -1 & 1 & 1 & -2 & 1 & 0 & 2 \\ 3 & -1 & 0 & 0 & 1 & 1 & 2 & -3 & 1 \\ -5 & 2 & -1 & 3 & 3 & -6 & 0 & -3 & 0 \end{bmatrix}.$$

The first three atoms are class 1, the second three class 2, the last three class 3.

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$.

To which class should be assign $\mathbf{x}^*$?

An Alternate Approach: Example 2

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$.

An Alternate Approach: Example 2

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$. To which class should be assign $\mathbf{x}^*$?

An Alternate Approach: Example 2

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$. To which class should be assign $\mathbf{x}^*$?

We form a sparse solution $\boldsymbol{\alpha}$ to the system $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$. The 3-sparse solution here is

$$\boldsymbol{\alpha} = \langle 0.267, 0, 1.1, 0, 0, 0, 0, -0.056, 0 \rangle.$$

An Alternate Approach: Example 2

Suppose we measure feature vector $\mathbf{x}^* = \langle 2.1, -1.1, 0.9, 2.4 \rangle$. To which class should be assign $\mathbf{x}^*$?

We form a sparse solution $\boldsymbol{\alpha}$ to the system $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$. The 3-sparse solution here is

$$\boldsymbol{\alpha} = \langle 0.267, 0, 1.1, 0, 0, 0, 0, -0.056, 0 \rangle.$$

Define class 1, 2, 3 portions

$$\begin{aligned}
\boldsymbol{\alpha}^1 &= \langle 0.267, 0, 1.1, 0, 0, 0, 0, 0, 0 \rangle \\
\boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, 0, 0, 0, 0, 0 \rangle \\
\boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, 0, 0, 0, -0.056, 0 \rangle.
\end{aligned}$$

Note $\boldsymbol{\alpha} = \boldsymbol{\alpha}^1 + \boldsymbol{\alpha}^2 + \boldsymbol{\alpha}^3$.

An Alternate Approach: Example 2

Note that

$$\begin{aligned} \boldsymbol{\alpha}^1 &= \langle 0.267, 0, 1.1, 0, 0, 0, 0, 0, 0\rangle \\ \boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, 0, 0, 0, 0, 0\rangle \\ \boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, 0, 0, 0, -0.056, 0\rangle. \end{aligned}$$

$\boldsymbol{\alpha}^1$ corresponds to atoms in class 1, $\boldsymbol{\alpha}^2$ to atoms in class 2,$\boldsymbol{\alpha}^3$ to atoms in class 3.

An Alternate Approach: Example 2

Note that

$$\boldsymbol{\alpha}^1 = \langle 0.267, 0, 1.1, 0, 0, 0, 0, 0, 0 \rangle$$
$$\boldsymbol{\alpha}^2 = \langle 0, 0, 0, 0, 0, 0, 0, 0, 0 \rangle$$
$$\boldsymbol{\alpha}^3 = \langle 0, 0, 0, 0, 0, 0, 0, -0.056, 0 \rangle.$$

$\boldsymbol{\alpha}^1$ corresponds to atoms in class 1, $\boldsymbol{\alpha}^2$ to atoms in class 2,$\boldsymbol{\alpha}^3$ to atoms in class 3.

The hope is that if $\mathbf{x}^*$ can be built sparsely and accurately as $\mathbf{D}\boldsymbol{\alpha}^1$ then $\mathbf{x}^*$ is in Class 1.

An Alternate Approach: Example 2

Note that

$$\begin{aligned}\boldsymbol{\alpha}^1 &= \langle 0.267, 0, 1.1, 0, 0, 0, 0, 0, 0\rangle\\ \boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, 0, 0, 0, 0, 0\rangle\\ \boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, 0, 0, 0, -0.056, 0\rangle.\end{aligned}$$

$\boldsymbol{\alpha}^1$ corresponds to atoms in class 1, $\boldsymbol{\alpha}^2$ to atoms in class 2,$\boldsymbol{\alpha}^3$ to atoms in class 3.

The hope is that if $\mathbf{x}^*$ can be built sparsely and accurately as $\mathbf{D}\boldsymbol{\alpha}^1$ then $\mathbf{x}^*$ is in Class 1.

Ditto for the other classes.

An Alternate Approach: Example 2

Compute

$$\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}^*\| = 0.465, \quad \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}^*\| = 3.49, \quad \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}^*\| = 3.67.$$

An Alternate Approach: Example 2

Compute

$$\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}^*\| = 0.465, \quad \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}^*\| = 3.49, \quad \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}^*\| = 3.67.$$

Since $\mathbf{x}^*$ can be built most accurately with a few vectors in class 1, we assign $\mathbf{x}^*$ to class 1.

The General Approach

We have K classes and m-dimensional feature vectors.

The General Approach

We have K classes and m-dimensional feature vectors.

We form a dictionary $\mathbf{D}$, an $m \times n$ matrix

$$\mathbf{D} = \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{v}_1^1 & \cdots & \mathbf{v}_1^{n_1} & \mathbf{v}_2^1 & \cdots & \mathbf{v}_2^{n_2} & \cdots & \mathbf{v}_K^1 & \cdots & \mathbf{v}_K^{n_K} \end{array} \right]$$

where each $\mathbf{v}_k^j \in \mathbb{R}^m$ and $n = n_1 + \cdots + n_K$. Usually $m << n$.

The General Approach

We have K classes and m-dimensional feature vectors.

We form a dictionary $\mathbf{D}$, an $m \times n$ matrix

$$\mathbf{D} = \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{v}_1^1 & \cdots & \mathbf{v}_1^{n_1} & \mathbf{v}_2^1 & \cdots & \mathbf{v}_2^{n_2} & \cdots & \mathbf{v}_K^1 & \cdots & \mathbf{v}_K^{n_K} \end{array} \right]$$

where each $\mathbf{v}_k^j \in \mathbb{R}^m$ and $n = n_1 + \cdots + n_K$. Usually $m << n$.

The vector (atom) $\mathbf{v}_k^j$ is in class k. The number of atoms n_k in any class may be large.

The General Approach

Using

$$\mathbf{D} = \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{v}_1^1 & \cdots & \mathbf{v}_1^{n_1} & \mathbf{v}_2^1 & \cdots & \mathbf{v}_2^{n_2} & \cdots & \mathbf{v}_K^1 & \cdots & \mathbf{v}_K^{n_K} \end{array} \right]$$

we find a sparse solution $\boldsymbol{\alpha} \in \mathbb{R}^n$ to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$.

The General Approach

Using

$$\mathbf{D} = \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{v}_1^1 & \cdots & \mathbf{v}_1^{n_1} & \mathbf{v}_2^1 & \cdots & \mathbf{v}_2^{n_2} & \cdots & \mathbf{v}_K^1 & \cdots & \mathbf{v}_K^{n_K} \end{array}\right]$$

we find a sparse solution $\boldsymbol{\alpha} \in \mathbb{R}^n$ to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$.

Let $\boldsymbol{\alpha}^k$ be the vector in $\mathbf{R}^n$ whose nonzero components are those of $\boldsymbol{\alpha}$ corresponding to atoms in class k.

The General Approach

Using

$$\mathbf{D} = \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{v}_1^1 & \cdots & \mathbf{v}_1^{n_1} & \mathbf{v}_2^1 & \cdots & \mathbf{v}_2^{n_2} & \cdots & \mathbf{v}_K^1 & \cdots & \mathbf{v}_K^{n_K} \end{array} \right]$$

we find a sparse solution $\boldsymbol{\alpha} \in \mathbb{R}^n$ to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$.

Let $\boldsymbol{\alpha}^k$ be the vector in $\mathbf{R}^n$ whose nonzero components are those of $\boldsymbol{\alpha}$ corresponding to atoms in class k.

Assign $\mathbf{x}^*$ to the class with minimum residual $\|\mathbf{D}\boldsymbol{\alpha}^k - \mathbf{x}^*\|$, indicating that $\mathbf{x}^*$ can be built most accurately using only a few atoms from Class k.

Insights

Imagine if the dictionary contained “all possible” feature vectors $\mathbf{x}^*$ that could occur.

Insights

Imagine if the dictionary contained "all possible" feature vectors $\mathbf{x}^*$ that could occur.

In this case $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$ contains a 1-sparse solution, namely $\boldsymbol{\alpha}$ has a single nonzero component α_j, where $\mathbf{x}^*$ (or a multiple thereof) appears as the jth column of $\mathbf{D}$; we could assign $\mathbf{x}^*$ to the corresponding class.

Insights

Imagine if the dictionary contained "all possible" feature vectors $\mathbf{x}^*$ that could occur.

In this case $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$ contains a 1-sparse solution, namely $\boldsymbol{\alpha}$ has a single nonzero component α_j, where $\mathbf{x}^*$ (or a multiple thereof) appears as the jth column of $\mathbf{D}$; we could assign $\mathbf{x}^*$ to the corresponding class.

If the dictionary $\mathbf{D}$ is merely large and $\mathbf{x}^*$ is in class k then it should be possible to build $\mathbf{x}^*$ accurately using only a few atoms from class k.

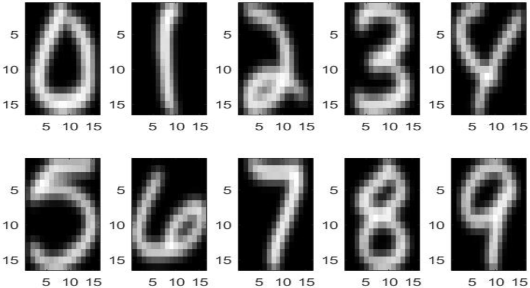
USPS Classification Example

The problem: classify handwritten digits, $0 - 9$.

USPS Classification Example

The problem: classify handwritten digits, $0 - 9$.

There are a total of 60000 digitized, handwritten digits in the MNIST database. Each is digitized on a 16×16 grid, normalized/registered:



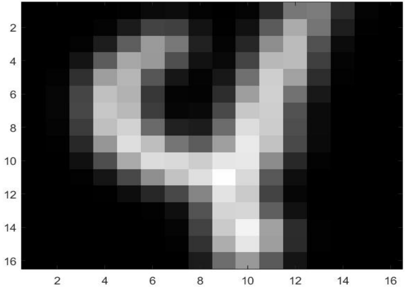
USPS Classification Example

We select 500 samples to act as the (initial) dictionary, about 50 samples of each digit. The additional set of $>$ 59000 handwritten digits can be used to test the classification algorithm.

USPS Classification Example

We select 500 samples to act as the (initial) dictionary, about 50 samples of each digit. The additional set of > 59000 handwritten digits can be used to test the classification algorithm.

A typical digit from this set:



Classification Example

The dictionary **D** consists of the 500 sample images, each arranged into a 256-dimensional vector, so **D** is a 256 by 500 matrix.

Classification Example

The dictionary $\mathbf{D}$ consists of the 500 sample images, each arranged into a 256-dimensional vector, so $\mathbf{D}$ is a 256 by 500 matrix.

To classify a signal $\mathbf{x}^*$ we

1. Run a sparse solver on $\mathbf{D}\alpha = \mathbf{x}^*$ (in this case, with sparsity limit 5).

Classification Example

The dictionary $\mathbf{D}$ consists of the 500 sample images, each arranged into a 256-dimensional vector, so $\mathbf{D}$ is a 256 by 500 matrix.

To classify a signal $\mathbf{x}^*$ we

1. Run a sparse solver on $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$ (in this case, with sparsity limit 5).
2. Form vectors $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^{10}$ ($\boldsymbol{\alpha}^j$ composed of the components of $\boldsymbol{\alpha}$ with support in class j.)

Classification Example

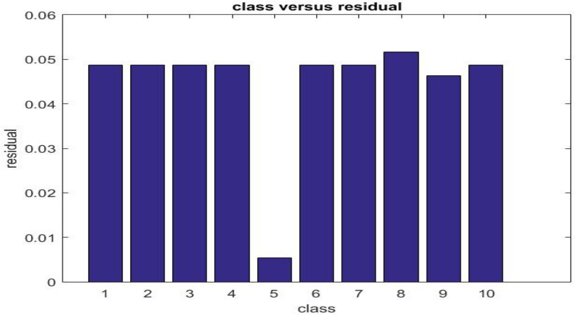
The dictionary $\mathbf{D}$ consists of the 500 sample images, each arranged into a 256-dimensional vector, so $\mathbf{D}$ is a 256 by 500 matrix.

To classify a signal $\mathbf{x}^*$ we

1. Run a sparse solver on $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}^*$ (in this case, with sparsity limit 5).
2. Form vectors $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^{10}$ ($\boldsymbol{\alpha}^j$ composed of the components of $\boldsymbol{\alpha}$ with support in class j.)
3. Assign $\mathbf{x}^*$ to the class k with minimum residual $\|\mathbf{D}\boldsymbol{\alpha}^k - \mathbf{x}\|$.

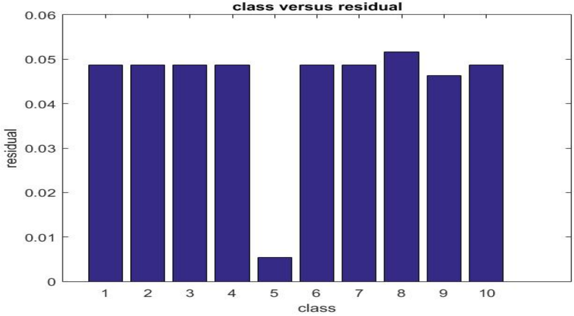
Classification Example

Residual $\|\mathbf{D}\alpha^k - \mathbf{x}^*\|_2$ for each class $k = 1$ through $k = 10$, for some handwritten digit image $\mathbf{x}^*$:



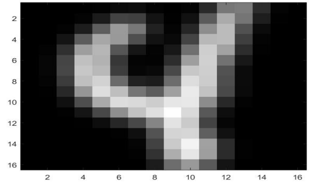
Classification Example

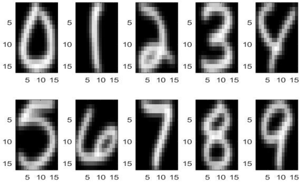
Residual $\|\mathbf{D}\alpha^k - \mathbf{x}^*\|_2$ for each class $k = 1$ through $k = 10$, for some handwritten digit image $\mathbf{x}^*$:

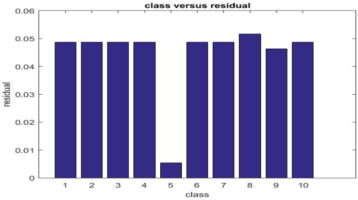


We conclude $\mathbf{x}^*$ is in class 5 (corresponding to a handwritten "4").

USPS Classification Example







Class 5 (digit "4") has the smallest residual.

Optimizing the Dictionary

If we use the additional 59500 samples to test our classification scheme, we obtain about 94 percent accuracy.

Optimizing the Dictionary

If we use the additional 59500 samples to test our classification scheme, we obtain about 94 percent accuracy.

But the initial dictionary is only a first step—we can improve the dictionary.

Optimizing the Dictionary

If we use the additional 59500 samples to test our classification scheme, we obtain about 94 percent accuracy.

But the initial dictionary is only a first step—we can improve the dictionary.

The first step is to select additional samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$; these will be used to "train" and improve the dictionary.

Optimizing the Dictionary: Example

Consider the previous example dictionary

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 2 & 1 & 2 & -1 & 1 & 3 & 1 \\ 1 & 0 & -1 & 1 & 1 & -2 & 1 & 0 & 2 \\ 3 & -1 & 0 & 0 & 1 & 1 & 2 & -3 & 1 \\ -5 & 2 & -1 & 3 & 3 & -6 & 0 & -3 & 0 \end{bmatrix}.$$

The first three atoms are class 1, the second three class 2, the last three class 3.

Optimizing the Dictionary: Example

Consider the previous example dictionary

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 2 & 1 & 2 & -1 & 1 & 3 & 1 \\ 1 & 0 & -1 & 1 & 1 & -2 & 1 & 0 & 2 \\ 3 & -1 & 0 & 0 & 1 & 1 & 2 & -3 & 1 \\ -5 & 2 & -1 & 3 & 3 & -6 & 0 & -3 & 0 \end{bmatrix}.$$

The first three atoms are class 1, the second three class 2, the last three class 3.

Suppose we have four additional samples, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, classes 1,2,3 and 3, respectively.

Optimizing the Dictionary: Example

Suppose $\mathbf{x}_1 = \langle 0.0.068, 1.31, 3.12, -5.11 \rangle$. Let's classify $\mathbf{x}_1$ as previously: Find a 3-sparse solution to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}_1$:

$$\boldsymbol{\alpha} = \langle 1.000, 0, 0, 0, -0.174, 0, 0, -0.144, 0 \rangle.$$

Optimizing the Dictionary: Example

Suppose $\mathbf{x}_1 = \langle 0.0.068, 1.31, 3.12, -5.11 \rangle$. Let's classify $\mathbf{x}_1$ as previously: Find a 3-sparse solution to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}_1$:

$$\boldsymbol{\alpha} = \langle 1.000, 0, 0, 0, -0.174, 0, 0, -0.144, 0 \rangle.$$

Define class 1/2/3 subsections

$$\begin{aligned} \boldsymbol{\alpha}^1 &= \langle 1.000, 0, 0, 0, 0, 0, 0, 0, 0 \rangle \\ \boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, -0.174, 0, 0, 0, 0 \rangle \\ \boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, -0, 0, 0, -0.144, 0 \rangle. \end{aligned}$$

Optimizing the Dictionary: Example

Suppose $\mathbf{x}_1 = \langle 0.0.068, 1.31, 3.12, -5.11\rangle$. Let's classify $\mathbf{x}_1$ as previously: Find a 3-sparse solution to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}_1$:

$$\boldsymbol{\alpha} = \langle 1.000, 0, 0, 0, -0.174, 0, 0, -0.144, 0\rangle.$$

Define class 1/2/3 subsections

$$\begin{aligned}\boldsymbol{\alpha}^1 &= \langle 1.000, 0, 0, 0, 0, 0, 0, 0, 0\rangle \\ \boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, -0.174, 0, 0, 0, 0\rangle \\ \boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, -0, 0, 0, -0.144, 0\rangle.\end{aligned}$$

We find $\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_1\|_2 = 0.996$, $\|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_1\|_2 = 5.85$, $\|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_1\|_2 = 6.31$, so $\mathbf{x}_1$ is assigned to Class 1 (correct).

Optimizing the Dictionary: Example

Suppose $\mathbf{x}_1 = \langle 0.0.068, 1.31, 3.12, -5.11\rangle$. Let's classify $\mathbf{x}_1$ as previously: Find a 3-sparse solution to $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}_1$:

$$\boldsymbol{\alpha} = \langle 1.000, 0, 0, 0, -0.174, 0, 0, -0.144, 0\rangle.$$

Define class 1/2/3 subsections

$$\begin{aligned}\boldsymbol{\alpha}^1 &= \langle 1.000, 0, 0, 0, 0, 0, 0, 0, 0\rangle\\ \boldsymbol{\alpha}^2 &= \langle 0, 0, 0, 0, -0.174, 0, 0, 0, 0\rangle\\ \boldsymbol{\alpha}^3 &= \langle 0, 0, 0, 0, -0, 0, 0, -0.144, 0\rangle.\end{aligned}$$

We find $\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_1\|_2 = 0.996$, $\|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_1\|_2 = 5.85$, $\|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_1\|_2 = 6.31$, so $\mathbf{x}_1$ is assigned to Class 1 (correct). But if the dictionary was perfect we'd have $\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_1\|_2 = 0$.

Optimizing the Dictionary: Example

A similar computation with class 2 training vector $\mathbf{x}_2 = \langle -0.380, 0.015, -0.510, 0.668 \rangle$ yields

$$\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_2\|_2 = 0.217, \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_2\|_2 = 0.922, \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_2\|_2 = 1.103.$$

Optimizing the Dictionary: Example

A similar computation with class 2 training vector $\mathbf{x}_2 = \langle -0.380, 0.015, -0.510, 0.668 \rangle$ yields

$$\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_2\|_2 = 0.217, \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_2\|_2 = 0.922, \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_2\|_2 = 1.103.$$

This training sample gets assigned to the wrong class (class 1 instead of class 2).

Optimizing the Dictionary: Example

A similar computation with class 2 training vector $\mathbf{x}_2 = \langle -0.380, 0.015, -0.510, 0.668 \rangle$ yields

$$\|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_2\|_2 = 0.217, \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_2\|_2 = 0.922, \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_2\|_2 = 1.103.$$

This training sample gets assigned to the wrong class (class 1 instead of class 2).

We'd like $\|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_2\|_2 = 0$.

Optimizing the Dictionary: Example

Training vectors $\mathbf{x}_3$ and $\mathbf{x}_4$ (not shown, but both class 3) yield residuals

$$\|\mathbf{D}\alpha^1 - \mathbf{x}_3\|_2 = 1.17, \quad \|\mathbf{D}\alpha^2 - \mathbf{x}_3\|_2 = 0.934, \quad \|\mathbf{D}\alpha^3 - \mathbf{x}_3\|_2 = 1.35$$
$$\|\mathbf{D}\alpha^1 - \mathbf{x}_4\|_2 = 3.12, \quad \|\mathbf{D}\alpha^2 - \mathbf{x}_4\|_2 = 3.39, \quad \|\mathbf{D}\alpha^3 - \mathbf{x}_4\|_2 = 0.638.$$

Optimizing the Dictionary: Example

Training vectors $\mathbf{x}_3$ and $\mathbf{x}_4$ (not shown, but both class 3) yield residuals

$$\|\mathbf{D}\alpha^1 - \mathbf{x}_3\|_2 = 1.17, \quad \|\mathbf{D}\alpha^2 - \mathbf{x}_3\|_2 = 0.934, \quad \|\mathbf{D}\alpha^3 - \mathbf{x}_3\|_2 = 1.35$$
$$\|\mathbf{D}\alpha^1 - \mathbf{x}_4\|_2 = 3.12, \quad \|\mathbf{D}\alpha^2 - \mathbf{x}_4\|_2 = 3.39, \quad \|\mathbf{D}\alpha^3 - \mathbf{x}_4\|_2 = 0.638.$$

A perfect dictionary would give $\|\mathbf{D}\alpha^3 - \mathbf{x}_3\|_2 = 0$ and $\|\mathbf{D}\alpha^3 - \mathbf{x}_4\|_2 = 0$.

Optimizing the Dictionary: Example

The training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_4$ are given, and for each $\mathbf{x}_i$ the sparse vector α is computed—think of them as now fixed.

Optimizing the Dictionary: Example

The training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_4$ are given, and for each $\mathbf{x}_i$ the sparse vector $\boldsymbol{\alpha}$ is computed—think of them as now fixed.

Consider the quantity

$$E(\mathbf{D}) = \|\mathbf{D}\boldsymbol{\alpha}^1 - \mathbf{x}_1\|_2^2 + \|\mathbf{D}\boldsymbol{\alpha}^2 - \mathbf{x}_2\|_2^2 + \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_3\|_2^2 + \|\mathbf{D}\boldsymbol{\alpha}^3 - \mathbf{x}_4\|_2^2$$

as a (quadratic) function of $\mathbf{D}$. A perfect dictionary would give $E(\mathbf{D}) = 0$.

Optimizing the Dictionary: Example

The training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_4$ are given, and for each $\mathbf{x}_i$ the sparse vector α is computed—think of them as now fixed.

Consider the quantity

$$E(\mathbf{D}) = \|\mathbf{D}\alpha^1 - \mathbf{x}_1\|_2^2 + \|\mathbf{D}\alpha^2 - \mathbf{x}_2\|_2^2 + \|\mathbf{D}\alpha^3 - \mathbf{x}_3\|_2^2 + \|\mathbf{D}\alpha^3 - \mathbf{x}_4\|_2^2$$

as a (quadratic) function of $\mathbf{D}$. A perfect dictionary would give $E(\mathbf{D}) = 0$.

We can improve the dictionary by minimizing $E(\mathbf{D})$ above, using any optimization algorithm.

Optimizing the Dictionary: Example

The training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_4$ are given, and for each $\mathbf{x}_i$ the sparse vector α is computed—think of them as now fixed.

Consider the quantity

$$E(\mathbf{D}) = \|\mathbf{D}\alpha^1 - \mathbf{x}_1\|_2^2 + \|\mathbf{D}\alpha^2 - \mathbf{x}_2\|_2^2 + \|\mathbf{D}\alpha^3 - \mathbf{x}_3\|_2^2 + \|\mathbf{D}\alpha^3 - \mathbf{x}_4\|_2^2$$

as a (quadratic) function of $\mathbf{D}$. A perfect dictionary would give $E(\mathbf{D}) = 0$.

We can improve the dictionary by minimizing $E(\mathbf{D})$ above, using any optimization algorithm.

We then go back and recompute each α for $\mathbf{x}_1, \ldots, \mathbf{x}_4$, and repeat the whole process.

Optimizing the Dictionary

Application of this process to the USPS dictionary increases the accuracy rate from 94 to 98 percent.

Optimizing the Dictionary

Application of this process to the USPS dictionary increases the accuracy rate from 94 to 98 percent.

Classification of any single handwritten digit takes about 1 millisecond (in Matlab).

Optimizing the Dictionary

Application of this process to the USPS dictionary increases the accuracy rate from 94 to 98 percent.

Classification of any single handwritten digit takes about 1 millisecond (in Matlab).

These techniques have been used in a wide variety of image classification problems, medical imaging, facial recognition, . . .