SVM multiclass classification in 10 steps

```
import numpy as np
# load digits dataset
from sklearn import datasets
digits = datasets.load_digits()
# define training set size
n_samples = len(digits.images)
n training = int(0.9 * n \text{ samples})
data = np.ascontiguousarray( digits.data, dtype=np.double )
labels = np.ascontiguousarrav( digits.target.reshape(n samples, 1).
                                dtvpe=np.double )
from daal.data_management import HomogenNumericTable
train_data = HomogenNumericTable( data[:n_training] )
train_labels = HomogenNumericTable( labels[:n_training] )
test data = HomogenNumericTable( data[n training:] )
```

SVM multiclass classification in 10 steps

```
import numpy as np
# load digits dataset
from sklearn import datasets
digits = datasets.load_digits()
# define training set size
n_samples = len(digits.images)
n training = int(0.9 * n \text{ samples})
data = np.ascontiguousarray( digits.data, dtype=np.double )
labels = np.ascontiguousarrav( digits.target.reshape(n samples, 1).
                                dtvpe=np.double )
from daal.data_management import HomogenNumericTable
train_data = HomogenNumericTable( data[:n_training] )
train_labels = HomogenNumericTable( labels[:n_training] )
test data = HomogenNumericTable( data[n training:] )
```

1. enjoy sklearn datasets import module

SVM multiclass classification in 10 steps

```
import numpy as np
# load digits dataset
from sklearn import datasets
digits = datasets.load_digits()
# define training set size
n_samples = len(digits.images)
n training = int(0.9 * n \text{ samples})
data = np.ascontiguousarray( digits.data, dtype=np.double )
labels = np.ascontiguousarrav( digits.target.reshape(n samples. 1).
                                dtvpe=np.double )
from daal.data_management import HomogenNumericTable
train_data = HomogenNumericTable( data[:n_training] )
train_labels = HomogenNumericTable( labels[:n_training] )
test data = HomogenNumericTable( data[n training:] )
```

- 1. enjoy sklearn datasets import module
- 2. require a contiguous array



SVM multiclass classification in 10 steps

```
import numpy as np
# load digits dataset
from sklearn import datasets
digits = datasets.load_digits()
# define training set size
n_samples = len(digits.images)
n training = int(0.9 * n \text{ samples})
data = np.ascontiguousarray( digits.data, dtype=np.double )
labels = np.ascontiguousarrav( digits.target.reshape(n samples, 1).
                                dtvpe=np.double )
from daal.data_management import HomogenNumericTable
train_data = HomogenNumericTable( data[:n_training] )
train_labels = HomogenNumericTable( labels[:n_training] )
test data = HomogenNumericTable( data[n training:] )
```

- 1. enjoy sklearn datasets import module
- 2. require a contiguous array
- 3. create instances of HomogenNumericTable



```
from daal.algorithms.svm import training as svm training
from daal.algorithms.svm import prediction as svm_prediction
from daal.algorithms.multi class classifier import training as
     multiclass training
from daal.algorithms.classifier import training as training_params
kernel = rbf.Batch Float64DefaultDense()
kernel.parameter.sigma = 0.001
# Create two class sym classifier
# training alg
twoclass_train_alg = svm_training.Batch_Float64DefaultDense()
twoclass train alg.parameter.kernel = kernel
twoclass train alg.parameter.C = 1.0
# prediction alg
twoclass_predict_alg = svm_prediction.Batch_Float64DefaultDense()
twoclass predict alg.parameter.kernel = kernel
# Create a multiclass classifier object (training)
train alg = multiclass training.Batch Float640neAgainstOne()
train alg.parameter.nClasses = 10
train_alg.parameter.training = twoclass_train_alg
train alg.parameter.prediction = twoclass predict alg
```

Cineca TRAINING High Performance Computing 2017

```
from daal.algorithms.svm import training as svm training
from daal.algorithms.svm import prediction as svm_prediction
from daal.algorithms.multi class classifier import training as
     multiclass training
from daal.algorithms.classifier import training as training params
kernel = rbf.Batch Float64DefaultDense()
kernel.parameter.sigma = 0.001
# Create two class sym classifier
# training alg
twoclass_train_alg = svm_training.Batch_Float64DefaultDense()
twoclass train alg.parameter.kernel = kernel
twoclass train alg.parameter.C = 1.0
# prediction alg
twoclass_predict_alg = svm_prediction.Batch_Float64DefaultDense()
twoclass predict alg.parameter.kernel = kernel
# Create a multiclass classifier object (training)
train alg = multiclass training.Batch Float640neAgainstOne()
train alg.parameter.nClasses = 10
train_alg.parameter.training = twoclass_train_alg
train alg.parameter.prediction = twoclass predict alg
```

4. define kernel and kernel parameters



```
from daal.algorithms.svm import training as svm training
from daal.algorithms.svm import prediction as svm_prediction
from daal.algorithms.multi class classifier import training as
     multiclass training
from daal.algorithms.classifier import training as training_params
kernel = rbf.Batch Float64DefaultDense()
kernel.parameter.sigma = 0.001
# Create two class sym classifier
# training alg
twoclass_train_alg = svm_training.Batch_Float64DefaultDense()
twoclass train alg.parameter.kernel = kernel
twoclass train alg.parameter.C = 1.0
# prediction alg
twoclass_predict_alg = svm_prediction.Batch_Float64DefaultDense()
twoclass_predict_alg.parameter.kernel = kernel
# Create a multiclass classifier object (training)
train alg = multiclass training.Batch Float640neAgainstOne()
train alg.parameter.nClasses = 10
train_alg.parameter.training = twoclass_train_alg
train alg.parameter.prediction = twoclass predict alg
```

- 4. define kernel and kernel parameters
- 5 create two class svm classifier



```
from daal.algorithms.svm import training as svm training
from daal.algorithms.svm import prediction as svm_prediction
from daal.algorithms.multi class classifier import training as
     multiclass training
from daal.algorithms.classifier import training as training_params
kernel = rbf.Batch Float64DefaultDense()
kernel.parameter.sigma = 0.001
# Create two class sym classifier
# training alg
twoclass_train_alg = svm_training.Batch_Float64DefaultDense()
twoclass train alg.parameter.kernel = kernel
twoclass train alg.parameter.C = 1.0
# prediction alg
twoclass_predict_alg = svm_prediction.Batch_Float64DefaultDense()
twoclass predict alg.parameter.kernel = kernel
# Create a multiclass classifier object (training)
train alg = multiclass training.Batch Float640neAgainstOne()
train alg.parameter.nClasses = 10
train_alg.parameter.training = twoclass_train_alg
train alg.parameter.prediction = twoclass predict alg
```

- 4. define kernel and kernel parameters
- 5 create two class svm classifier
- 6. create multi class svm classifier (training)





```
# Pass training data and labels
train_alg.input.set(training_params.data, train_data)
train_alg.input.set(training_params.labels, train_labels)
# training
model = train_alg.compute().get(training_params.model)
```







(# Pass training data and labels train_alg.input.set(training_params.data, train_data) train_alg.input.set(training_params.labels, train_labels)

training
model = train_alg.compute().get(training_params.model)

7. set input data and labels





- 7. set input data and labels
- 8. start training and get model



prediction algorithm setup

99111

```
from daal.algorithms.multi_class_classifier import prediction as
    multiclass_prediction
from daal.algorithms.classifier import prediction as
    prediction_params
# Create a multiclass classifier object (prediction)
predict_alg = multiclass_prediction.
    Batch_Float64DefaultDenseOneAgainstOne()
predict_alg.parameter.nClasses = 10
predict_alg.parameter.training = twoclass_train_alg
predict_alg.parameter.training = twoclass_predict_alg
```

Cineca

High Performance

prediction algorithm setup

Cineca TRAINING High Performance Computing 2017



9. create multi class svm classifier (prediction)





10. set input model and data





- 10. set input model and data
- 11. start prediction and get labels



Benchmark results

- Test description: 1797 samples total (90% training set, 10% test set), 64 features per sample
- Platform description: Intel Core i5-6300U CPU @ 2.40GHz

	sklearn	pyDAAL	speedup
training time [s]	0.161	0.018	8.9
test time [s]	0.017	0.004	4.3



Cineca

High Performance

DAAL in general:





DAAL in general:

✓ very simple installation/setup



Cineca

High Performance

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)



Cineca

High Performance

DAAL in general:

- very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)

Cineca

High Performance

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)
- X documentation is sometimes not exhaustive

Cineca TRAINING High Performance Computing 2017

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)
- X documentation is sometimes not exhaustive
- × examples cover very simple application cases

Cineca TRAINING High Performance Computing 2017

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)
- X documentation is sometimes not exhaustive
- × examples cover very simple application cases

as a Python user:

✓ comes with Intel Python framework



Cineca TRAINING High Performance Computing 2017

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)
- X documentation is sometimes not exhaustive
- × examples cover very simple application cases

as a Python user:

- ✓ comes with Intel Python framework
- ✓ faster alternative to scikit

Cineca TRAINING High Performance Computing 2017

DAAL in general:

- ✓ very simple installation/setup
- ✓ wide range of algorithms (both for machine l. and for deep l.)
- ✓ good support through dedicated intel forum (even for non paid versions)
- ✓ DAAL C⁺⁺ can be called from R and Matlab (see how-to forum posts)
- X documentation is sometimes not exhaustive
- × examples cover very simple application cases

as a Python user:

- ✓ comes with Intel Python framework
- ✓ faster alternative to scikit
- X Python interface still in development phase
 - not all neural network layers parameters are accessible/modifiable

NVIDIA Deep Learning Software Cineca TRAINING High Performance Computing 2017 uDNN TensorRT UDNN TensorRT UDNN TensorRT UDNN TensorRT UDNN TensorRT UDIT TensorRT



Cuber Cineca NVIDIA Deep Learning Software Cineca Computing 2017

Tools and libraries for designing and deploying GPU-accelerated deep learning applications.

 Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;





- Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;
- TensorRT optimize, validate and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive product platforms;

Cineca TRAINING High Performance Computing 2017



- Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;
- TensorRT optimize, validate and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive product platforms;
- DeepStream SDK uses TensorRT to deliver fast INT8 precision inference for real-time video content analysis, supports also FP16 and FP32 precisions;

Cineca TRAINING High Performance Computing 2017



- Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;
- TensorRT optimize, validate and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive product platforms;
- DeepStream SDK uses TensorRT to deliver fast INT8 precision inference for real-time video content analysis, supports also FP16 and FP32 precisions;
- Linear Algebra (cuBLAS and cuBLAS-XT) accelerated BLAS subroutines for single and multi-GPU acceleration;

Cineca TRAINING High Performance Computing 2017



- Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;
- TensorRT optimize, validate and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive product platforms;
- DeepStream SDK uses TensorRT to deliver fast INT8 precision inference for real-time video content analysis, supports also FP16 and FP32 precisions;
- Linear Algebra (cuBLAS and cuBLAS-XT) accelerated BLAS subroutines for single and multi-GPU acceleration;
- Sparse Linear Algebra (cuSPARSE) supports dense, COO, CSR, CSC, ELL/HYB and Blocked CSR sparse matrix formats, Level 1,2,3 routines, sparse triangular solver, sparse tridiagonal solver;

Cineca TRAINING High Performance Computing 2017



- Deep Learning Neural Network library (cuDNN) forward and backward convolution, pooling, normalization, activation layers;
- TensorRT optimize, validate and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive product platforms;
- DeepStream SDK uses TensorRT to deliver fast INT8 precision inference for real-time video content analysis, supports also FP16 and FP32 precisions;
- Linear Algebra (cuBLAS and cuBLAS-XT) accelerated BLAS subroutines for single and multi-GPU acceleration;
- Sparse Linear Algebra (cuSPARSE) supports dense, COO, CSR, CSC, ELL/HYB and Blocked CSR sparse matrix formats, Level 1,2,3 routines, sparse triangular solver, sparse tridiagonal solver;
- Multi-GPU Communications (NCCL, pronounced "Nickel") optimized primitives for collective multi-GPU communication;



NVIDIA based frameworks





Cineca TRAINING High Performance Computing 2017

• Google Brain's second generation machine learning system



- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs

- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs
- automatic differentiation capabilities

- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs
- automatic differentiation capabilities
- optimization algorithms: gradient and proximal gradient based

- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs
- automatic differentiation capabilities
- optimization algorithms: gradient and proximal gradient based
- code portability (CPUs, GPUs, on desktop, server, or mobile computing platforms)

- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs
- automatic differentiation capabilities
- optimization algorithms: gradient and proximal gradient based
- code portability (CPUs, GPUs, on desktop, server, or mobile computing platforms)
- Python interface is the preferred one (Java and C++also exist)

- Google Brain's second generation machine learning system
- computations are expressed as stateful dataflow graphs
- automatic differentiation capabilities
- optimization algorithms: gradient and proximal gradient based
- code portability (CPUs, GPUs, on desktop, server, or mobile computing platforms)
- Python interface is the preferred one (Java and C++also exist)
- installation through: virtualenv, pip, Docker, Anaconda, from sources



1. generate noisy input data

```
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
# Set up the data with a noisy linear relationship between X and Y.
num_examples = 50
X = np.array([np.linspace(-2, 4, num_examples), np.linspace(-6, 6,
num_examples])
X += np.random.randn(2, num_examples)
x, y = X
x_with_bias = np.array([(1., a) for a in x]).astype(np.float32)
losses = []
training_steps = 50
learning_rate = 0.002
```



- 1. generate noisy input data
- 2. set slack variables and fix algorithm parameters

Cineca

High Performance

99111

```
# Start of graph description
# Set up all the tensors, variables, and operations.
A = tf.constant(x_with_bias)
target = tf.constant(np.transpose([v]).astype(np.float32))
weights = tf.Variable(tf.random normal([2, 1], 0, 0.1))
yhat = tf.matmul(A, weights)
verror = tf.sub(vhat, target)
loss = tf.nn.12_loss(yerror)
update_weights =
  tf.train.GradientDescentOptimizer(learning_rate).minimize(loss)
sess = tf.Session()
sess.run( tf.global variables initializer() )
for _ in range(training_steps):
  # Repeatedly run the operations, updating variables
  sess.run( update weights )
  losses.append( sess.run( loss ) )
```

TRAINING High Performance Computing 2017

Cineca

```
# Start of graph description
# Set up all the tensors, variables, and operations.
A = tf.constant(x_with_bias)
target = tf.constant(np.transpose([v]).astype(np.float32))
weights = tf.Variable(tf.random normal([2, 1], 0, 0.1))
yhat = tf.matmul(A, weights)
verror = tf.sub(vhat, target)
loss = tf.nn.12_loss(yerror)
update_weights =
  tf.train.GradientDescentOptimizer(learning_rate).minimize(loss)
sess = tf.Session()
sess.run( tf.global variables initializer() )
for _ in range(training_steps):
  # Repeatedly run the operations, updating variables
  sess.run( update weights )
  losses.append( sess.run( loss ) )
```

3. define tensorflow constants and variables



Cineca

High Performance

```
# Start of graph description
# Set up all the tensors, variables, and operations.
A = tf.constant(x_with_bias)
target = tf.constant(np.transpose([v]).astype(np.float32))
weights = tf.Variable(tf.random normal([2, 1], 0, 0.1))
yhat = tf.matmul(A, weights)
verror = tf.sub(vhat, target)
loss = tf.nn.12_loss(yerror)
update_weights =
  tf.train.GradientDescentOptimizer(learning_rate).minimize(loss)
sess = tf.Session()
sess.run( tf.global variables initializer() )
for _ in range(training_steps):
  # Repeatedly run the operations, updating variables
  sess.run( update weights )
  losses.append( sess.run( loss ) )
```

- 3. define tensorflow constants and variables
- 4. define nodes

Cineca

High Performance

```
# Start of graph description
# Set up all the tensors, variables, and operations.
A = tf.constant(x_with_bias)
target = tf.constant(np.transpose([v]).astype(np.float32))
weights = tf.Variable(tf.random normal([2, 1], 0, 0.1))
yhat = tf.matmul(A, weights)
verror = tf.sub(vhat, target)
loss = tf.nn.12_loss(yerror)
update_weights =
  tf.train.GradientDescentOptimizer(learning_rate).minimize(loss)
sess = tf.Session()
sess.run( tf.global variables initializer() )
for _ in range(training_steps):
  # Repeatedly run the operations, updating variables
  sess.run( update weights )
  losses.append( sess.run( loss ) )
```

- 3. define tensorflow constants and variables
- 4. define nodes
- 5. start evaluation



Cineca

High Performance

Cineca TRAINING High Performance Computing 2017

Training is done, get the final values betas = sess.run(weights) yhat = sess.run(yhat)



99444

Cineca TRAINING High Performance Computing 2017

Training is done, get the final values betas = sess.run(weights) yhat = sess.run(yhat)



Benchmark results

- MNIST dataset of handwritten digits:
 - training set: 60k samples, 784 features per sample (MNIST)
 - test set: 10k samples, 784 features per sample (MNIST)



Benchmark results

99111

- MNIST dataset of handwritten digits:
 - training set: 60k samples, 784 features per sample (MNIST)
 - test set: 10k samples, 784 features per sample (MNIST)
- Convolutional NN: two conv. layers, two fully conn. layers (plus reg.) \approx 3m variables



Optimization method:

- stochastic gradient descent (batch size: 50 examples)
- fixed learning rate
- 2000 iterations



Cineca

High Performance

Optimization method:

- stochastic gradient descent (batch size: 50 examples)
- fixed learning rate
- 2000 iterations

Platforms:

- (1) Intel Core i5-6300U CPU @2.4GHz
- (2) 2 x Intel Xeon 2630 v3 @2.4GHz
- (3) Nvidia Tesla K40



Optimization method:

- stochastic gradient descent (batch size: 50 examples)
- fixed learning rate
- 2000 iterations

Platforms:

- (1) Intel Core i5-6300U CPU @2.4GHz
- (2) 2 x Intel Xeon 2630 v3 @2.4GHz
- (3) Nvidia Tesla K40

	PL (1)	PL (2)	PL (3)
training time [s]	3307.2	866.1	191.8
test time [s]	11.9	1.7	1.2



Optimization method:

- stochastic gradient descent (batch size: 50 examples)
- fixed learning rate
- 2000 iterations

Platforms:

- (1) Intel Core i5-6300U CPU @2.4GHz
- (2) 2 x Intel Xeon 2630 v3 @2.4GHz
- (3) Nvidia Tesla K40

	PL (1)	PL (2)	PL (3)
training time [s]	3307.2	866.1	191.8
test time [s]	11.9	1.7	1.2

Same code achieves 3.8x when running in 1 GALILEO node and 17.2x on a single GPU (training phase).



Cineca

High Performance

Cineca TRAINING High Performance

Computing 2017

TensorFlow in general:



Cineca TRAINING High Performance Computing 2017

TensorFlow in general:

✓ very simple installation/setup



Cineca TRAINING High Performance Computing 2017

TensorFlow in general:

- ✓ very simple installation/setup
- \checkmark plenty of tutorials, exhaustive documentation

TensorFlow in general:

- very simple installation/setup
- \checkmark plenty of tutorials, exhaustive documentation
- ✓ tools for exporting (partially) trained graphs (see MetaGraph)



Cineca

High Performance

TensorFlow in general:

- ✓ very simple installation/setup
- ✓ plenty of tutorials, exhaustive documentation
- ✓ tools for exporting (partially) trained graphs (see MetaGraph)
- ✓ debugger, graph flow visualization

Cineca

High Performance

TensorFlow in general:

- very simple installation/setup
- ✓ plenty of tutorials, exhaustive documentation
- ✓ tools for exporting (partially) trained graphs (see MetaGraph)
- ✓ debugger, graph flow visualization

as a Python user:

✓ "Python API is the most complete and the easiest to use"

Cineca

High Performance

TensorFlow in general:

- very simple installation/setup
- ✓ plenty of tutorials, exhaustive documentation
- ✓ tools for exporting (partially) trained graphs (see MetaGraph)
- ✓ debugger, graph flow visualization

as a Python user:

- ✓ "Python API is the most complete and the easiest to use"
- ✓ numpy interoperability

Cineca

High Performance

TensorFlow in general:

- very simple installation/setup
- \checkmark plenty of tutorials, exhaustive documentation
- ✓ tools for exporting (partially) trained graphs (see MetaGraph)
- ✓ debugger, graph flow visualization

as a Python user:

- "Python API is the most complete and the easiest to use"
- ✓ numpy interoperability
- ✗ lower level than pyDAAL (?)



Cineca

High Performance