**1.5. Stochastic Gradient Descent**

**Stochastic Gradient Descent (SGD)** is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) [Support Vector Machines](https://en.wikipedia.org/wiki/Support_vector_machine) and [Logistic Regression](https://en.wikipedia.org/wiki/Logistic_regression). Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning.

SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features.

The advantages of Stochastic Gradient Descent are:

* Efficiency.
* Ease of implementation (lots of opportunities for code tuning).

The disadvantages of Stochastic Gradient Descent include:

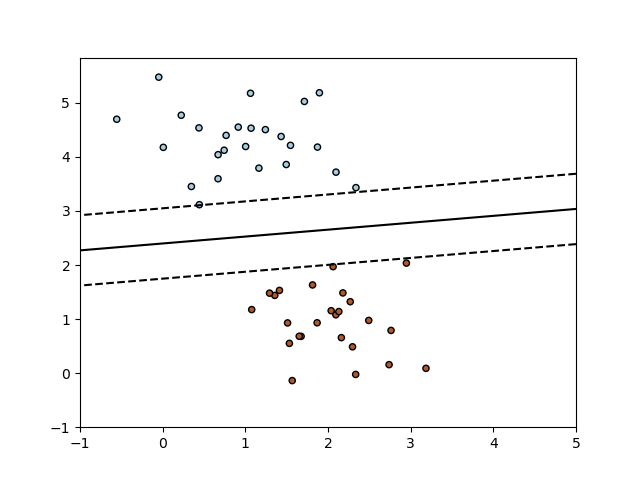
* SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations.
* SGD is sensitive to feature scaling.

**1.5.1. Classification**

**Warning**

Make sure you permute (shuffle) your training data before fitting the model or use shuffle=True to shuffle after each iteration.

The class [**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification.

[](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_separating_hyperplane.html)

As other classifiers, SGD has to be fitted with two arrays: an array X of size [n\_samples, n\_features] holding the training samples, and an array Y of size [n\_samples] holding the target values (class labels) for the training samples:

>>>

**>>> from** **sklearn.linear\_model** **import** SGDClassifier

**>>>** X = [[0., 0.], [1., 1.]]

**>>>** y = [0, 1]

**>>>** clf = SGDClassifier(loss="hinge", penalty="l2")

**>>>** clf.fit(X, y)

SGDClassifier(alpha=0.0001, average=False, class\_weight=None, epsilon=0.1,

eta0=0.0, fit\_intercept=True, l1\_ratio=0.15,

learning\_rate='optimal', loss='hinge', max\_iter=None, n\_iter=None,

n\_jobs=1, penalty='l2', power\_t=0.5, random\_state=None,

shuffle=True, tol=None, verbose=0, warm\_start=False)

After being fitted, the model can then be used to predict new values:

>>>

**>>>** clf.predict([[2., 2.]])

array([1])

SGD fits a linear model to the training data. The member coef\_ holds the model parameters:

>>>

**>>>** clf.coef\_

array([[ 9.9..., 9.9...]])

Member intercept\_ holds the intercept (aka offset or bias):

>>>

**>>>** clf.intercept\_

array([-9.9...])

Whether or not the model should use an intercept, i.e. a biased hyperplane, is controlled by the parameter fit\_intercept.

To get the signed distance to the hyperplane use [**SGDClassifier.decision\_function**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier.decision_function):

>>>

**>>>** clf.decision\_function([[2., 2.]])

array([ 29.6...])

The concrete loss function can be set via the loss parameter. [**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) supports the following loss functions:

* loss="hinge": (soft-margin) linear Support Vector Machine,
* loss="modified\_huber": smoothed hinge loss,
* loss="log": logistic regression,
* and all regression losses below.

The first two loss functions are lazy, they only update the model parameters if an example violates the margin constraint, which makes training very efficient and may result in sparser models, even when L2 penalty is used.

Using loss="log" or loss="modified\_huber" enables the predict\_proba method, which gives a vector of probability estimates P(y|x) per sample x:

>>>

**>>>** clf = SGDClassifier(loss="log").fit(X, y)

**>>>** clf.predict\_proba([[1., 1.]])

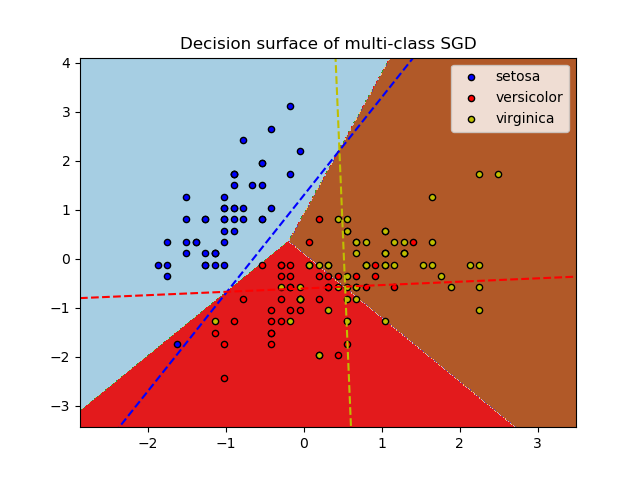
array([[ 0.00..., 0.99...]])

The concrete penalty can be set via the penalty parameter. SGD supports the following penalties:

* penalty="l2": L2 norm penalty on coef\_.
* penalty="l1": L1 norm penalty on coef\_.
* penalty="elasticnet": Convex combination of L2 and L1; (1 - l1\_ratio) \* L2 + l1\_ratio \* L1.

The default setting is penalty="l2". The L1 penalty leads to sparse solutions, driving most coefficients to zero. The Elastic Net solves some deficiencies of the L1 penalty in the presence of highly correlated attributes. The parameter l1\_ratiocontrols the convex combination of L1 and L2 penalty.

[**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) supports multi-class classification by combining multiple binary classifiers in a “one versus all” (OVA) scheme. For each of the K classes, a binary classifier is learned that discriminates between that and all other K-1classes. At testing time, we compute the confidence score (i.e. the signed distances to the hyperplane) for each classifier and choose the class with the highest confidence. The Figure below illustrates the OVA approach on the iris dataset. The dashed lines represent the three OVA classifiers; the background colors show the decision surface induced by the three classifiers.

[](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_iris.html)

In the case of multi-class classification coef\_ is a two-dimensionally array of shape=[n\_classes, n\_features] and intercept\_ is a one dimensional array of shape=[n\_classes]. The i-th row of coef\_ holds the weight vector of the OVA classifier for the i-th class; classes are indexed in ascending order (see attribute classes\_). Note that, in principle, since they allow to create a probability model, loss="log" and loss="modified\_huber" are more suitable for one-vs-all classification.

[**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) supports both weighted classes and weighted instances via the fit parameters class\_weight and sample\_weight. See the examples below and the doc string of [**SGDClassifier.fit**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier.fit) for further information.

**Examples:**

* [SGD: Maximum margin separating hyperplane](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_separating_hyperplane.html#sphx-glr-auto-examples-linear-model-plot-sgd-separating-hyperplane-py),
* [Plot multi-class SGD on the iris dataset](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_iris.html#sphx-glr-auto-examples-linear-model-plot-sgd-iris-py)
* [SGD: Weighted samples](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_weighted_samples.html#sphx-glr-auto-examples-linear-model-plot-sgd-weighted-samples-py)
* [Comparing various online solvers](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_comparison.html#sphx-glr-auto-examples-linear-model-plot-sgd-comparison-py)
* [SVM: Separating hyperplane for unbalanced classes](http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane_unbalanced.html#sphx-glr-auto-examples-svm-plot-separating-hyperplane-unbalanced-py) (See the Note)

[**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) supports averaged SGD (ASGD). Averaging can be enabled by setting `average=True`. ASGD works by averaging the coefficients of the plain SGD over each iteration over a sample. When using ASGD the learning rate can be larger and even constant leading on some datasets to a speed up in training time.

For classification with a logistic loss, another variant of SGD with an averaging strategy is available with Stochastic Average Gradient (SAG) algorithm, available as a solver in [**LogisticRegression**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression).

**1.5.2. Regression**

The class [**SGDRegressor**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor) implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties to fit linear regression models. [**SGDRegressor**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor) is well suited for regression problems with a large number of training samples (> 10.000), for other problems we recommend [**Ridge**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge), [**Lasso**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso), or [**ElasticNet**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html#sklearn.linear_model.ElasticNet).

The concrete loss function can be set via the loss parameter. [**SGDRegressor**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor) supports the following loss functions:

* loss="squared\_loss": Ordinary least squares,
* loss="huber": Huber loss for robust regression,
* loss="epsilon\_insensitive": linear Support Vector Regression.

The Huber and epsilon-insensitive loss functions can be used for robust regression. The width of the insensitive region has to be specified via the parameter epsilon. This parameter depends on the scale of the target variables.

[**SGDRegressor**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor) supports averaged SGD as [**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier). Averaging can be enabled by setting `average=True`.

For regression with a squared loss and a l2 penalty, another variant of SGD with an averaging strategy is available with Stochastic Average Gradient (SAG) algorithm, available as a solver in [**Ridge**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge).

**1.5.3. Stochastic Gradient Descent for sparse data**

**Note**

The sparse implementation produces slightly different results than the dense implementation due to a shrunk learning rate for the intercept.

There is built-in support for sparse data given in any matrix in a format supported by [scipy.sparse](https://docs.scipy.org/doc/scipy/reference/sparse.html). For maximum efficiency, however, use the CSR matrix format as defined in [scipy.sparse.csr\_matrix](http://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html).

**Examples:**

* [Classification of text documents using sparse features](http://scikit-learn.org/stable/auto_examples/text/document_classification_20newsgroups.html#sphx-glr-auto-examples-text-document-classification-20newsgroups-py)

**1.5.4. Complexity**

The major advantage of SGD is its efficiency, which is basically linear in the number of training examples. If X is a matrix of size (n, p) training has a cost of O(k n \bar p), where k is the number of iterations (epochs) and \bar p is the average number of non-zero attributes per sample.

Recent theoretical results, however, show that the runtime to get some desired optimization accuracy does not increase as the training set size increases.

**1.5.5. Tips on Practical Use**

* Stochastic Gradient Descent is sensitive to feature scaling, so it is highly recommended to scale your data. For example, scale each attribute on the input vector X to [0,1] or [-1,+1], or standardize it to have mean 0 and variance 1. Note that the *same* scaling must be applied to the test vector to obtain meaningful results. This can be easily done using **StandardScaler**:
* **from** **sklearn.preprocessing** **import** StandardScaler
* scaler = StandardScaler()
* scaler.fit(X\_train) *# Don't cheat - fit only on training data*
* X\_train = scaler.transform(X\_train)
* X\_test = scaler.transform(X\_test) *# apply same transformation to test data*

If your attributes have an intrinsic scale (e.g. word frequencies or indicator features) scaling is not needed.

* Finding a reasonable regularization term \alpha is best done using **GridSearchCV**, usually in the range 10.0\*\*-np.arange(1,7).
* Empirically, we found that SGD converges after observing approx. 10^6 training samples. Thus, a reasonable first guess for the number of iterations is n\_iter = np.ceil(10\*\*6 / n), where n is the size of the training set.
* If you apply SGD to features extracted using PCA we found that it is often wise to scale the feature values by some constant c such that the average L2 norm of the training data equals one.
* We found that Averaged SGD works best with a larger number of features and a higher eta0

**References:**

* [“Efficient BackProp”](http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf) Y. LeCun, L. Bottou, G. Orr, K. Müller - In Neural Networks: Tricks of the Trade 1998.

**1.5.6. Mathematical formulation**

Given a set of training examples (x_1, y_1), \ldots, (x_n, y_n) where x_i \in \mathbf{R}^m and y_i \in \{-1,1\}, our goal is to learn a linear scoring function f(x) = w^T x + b with model parameters w \in \mathbf{R}^m and intercept b \in \mathbf{R}. In order to make predictions, we simply look at the sign of f(x). A common choice to find the model parameters is by minimizing the regularized training error given by

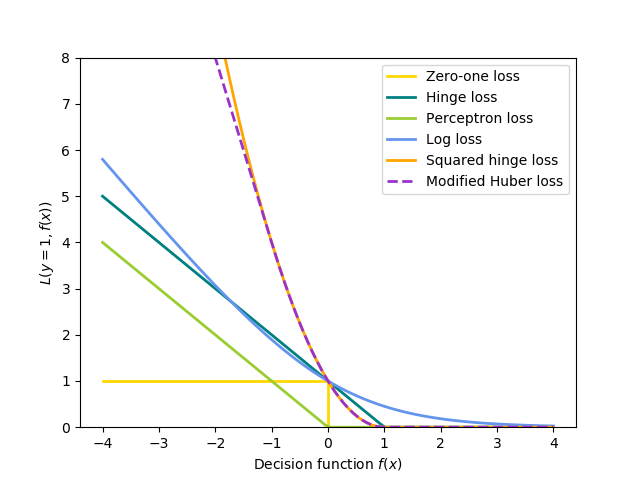
E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w)

where L is a loss function that measures model (mis)fit and R is a regularization term (aka penalty) that penalizes model complexity; \alpha > 0 is a non-negative hyperparameter.

Different choices for L entail different classifiers such as

* Hinge: (soft-margin) Support Vector Machines.
* Log: Logistic Regression.
* Least-Squares: Ridge Regression.
* Epsilon-Insensitive: (soft-margin) Support Vector Regression.

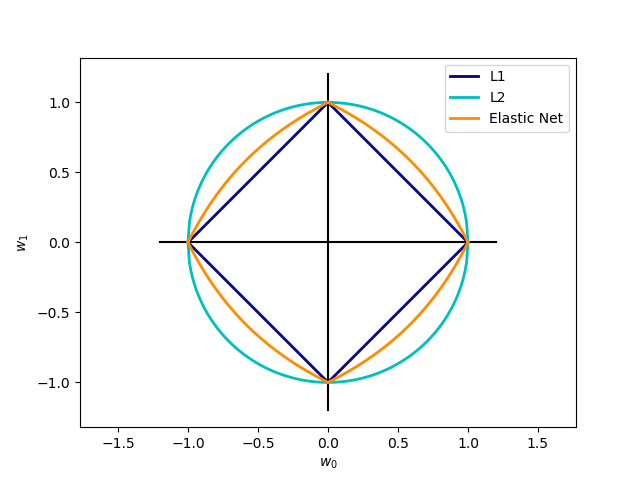
All of the above loss functions can be regarded as an upper bound on the misclassification error (Zero-one loss) as shown in the Figure below.

[](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_loss_functions.html)

Popular choices for the regularization term R include:

* L2 norm: R(w) := \frac{1}{2} \sum_{i=1}^{n} w_i^2,
* L1 norm: R(w) := \sum_{i=1}^{n} |w_i|, which leads to sparse solutions.
* Elastic Net: R(w) := \frac{\rho}{2} \sum_{i=1}^{n} w_i^2 + (1-\rho) \sum_{i=1}^{n} |w_i|, a convex combination of L2 and L1, where \rho is given by 1 - l1\_ratio.

The Figure below shows the contours of the different regularization terms in the parameter space when R(w) = 1.

[](http://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_penalties.html)

**1.5.6.1. SGD**

Stochastic gradient descent is an optimization method for unconstrained optimization problems. In contrast to (batch) gradient descent, SGD approximates the true gradient of E(w,b) by considering a single training example at a time.

The class [**SGDClassifier**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier) implements a first-order SGD learning routine. The algorithm iterates over the training examples and for each example updates the model parameters according to the update rule given by

w \leftarrow w - \eta (\alpha \frac{\partial R(w)}{\partial w}
+ \frac{\partial L(w^T x_i + b, y_i)}{\partial w})

where \eta is the learning rate which controls the step-size in the parameter space. The intercept b is updated similarly but without regularization.

The learning rate \eta can be either constant or gradually decaying. For classification, the default learning rate schedule (learning\_rate='optimal') is given by

\eta^{(t)} = \frac {1}{\alpha  (t_0 + t)}

where t is the time step (there are a total of n\_samples \* n\_iter time steps), t_0 is determined based on a heuristic proposed by Léon Bottou such that the expected initial updates are comparable with the expected size of the weights (this assuming that the norm of the training samples is approx. 1). The exact definition can be found in \_init\_t in **BaseSGD**.

For regression the default learning rate schedule is inverse scaling (learning\_rate='invscaling'), given by

\eta^{(t)} = \frac{eta_0}{t^{power\_t}}

where eta_0 and power\_t are hyperparameters chosen by the user via eta0 and power\_t, resp.

For a constant learning rate use learning\_rate='constant' and use eta0 to specify the learning rate.

The model parameters can be accessed through the members coef\_ and intercept\_:

* Member coef\_ holds the weights w
* Member intercept\_ holds b

**References:**

* [“Solving large scale linear prediction problems using stochastic gradient descent algorithms”](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.7377) T. Zhang - In Proceedings of ICML ‘04.
* [“Regularization and variable selection via the elastic net”](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.124.4696) H. Zou, T. Hastie - Journal of the Royal Statistical Society Series B, 67 (2), 301-320.
* [“Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent”](http://arxiv.org/pdf/1107.2490v2.pdf) Xu, Wei

**1.5.7. Implementation details**

The implementation of SGD is influenced by the [Stochastic Gradient SVM](http://leon.bottou.org/projects/sgd) of Léon Bottou. Similar to SvmSGD, the weight vector is represented as the product of a scalar and a vector which allows an efficient weight update in the case of L2 regularization. In the case of sparse feature vectors, the intercept is updated with a smaller learning rate (multiplied by 0.01) to account for the fact that it is updated more frequently. Training examples are picked up sequentially and the learning rate is lowered after each observed example. We adopted the learning rate schedule from Shalev-Shwartz et al. 2007. For multi-class classification, a “one versus all” approach is used. We use the truncated gradient algorithm proposed by Tsuruoka et al. 2009 for L1 regularization (and the Elastic Net). The code is written in Cython.