Finance Data Science
Lecture 7: Classification

Laurent El Ghaoui

MFE 230P, Summer 2017
MFE Program
Haas School of Business
UC Berkeley

6/26/2017

Notes

Outline

Notes

Basics of binary classification

Data

We are given a *training* data set with n measurements:

- *Feature vectors:* data points $x_i \in \mathbf{R}^p$, $i = 1, \ldots, n$.
- *Labels:* $y_i \in \{-1, 1\}$, $i = 1, \ldots, n$.

Examples:

Feature vectors	Labels
Companies' corporate info	default/no default
Stock price data	price up/down
News data	price up/down
News data	sentiment (positive/negative)
Emails	presence of a keyword
Genetic measures	presence of disease

Using the training data set $\{x_i, y_i\}_{i=1}^n$, our goal is to find a classification rule $\hat{y} = f(x)$ allowing to predict the label $\hat{y}$ of a new data point x.

Notes

Popular classification algorithms

- Naïve Bayes classifier;
- Support vector machines;
- Logistic regression;
- Decision trees and random forests;
- Neural networks;
- Etc.

In this lecture, we focus on SVM and logistic regression.

Finance Data Science
7. Classification
MFE 230P, Summer 2017
Binary Classification
Basics
SVM
Logistic regression
Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness
References

Notes

Linear classification

Linear classification rule: assumes f is a combination of the sign function and a linear (in fact, affine) function:

$$\hat{y} = \mathbf{sign}(w^T x + b),$$

where $w \in \mathbf{R}^p$, $b \in \mathbf{R}$ are given.

The goal of a linear classification algorithm is to find w, b, using the training data.

Notes

Multi-class problems

Finance Data Science
7. Classification

MFE 230P, Summer 2017

Binary Classification
Basics
SVM
Logistic regression
Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness
References

In some problems, the '"labels" y_i, $i = 1, \ldots, m$ are not binary, but correspond to more than two categories (*e.g.*, star ratings, analysts recommendations, etc).

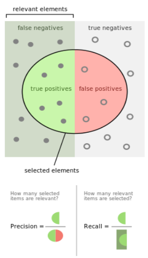
- A common practice is to transform the problem into a sequence of binary classification problems, doing multiple "one-vs-all" approaches.
- Some of the approaches discussed later can handle directly multi-class problems.
- If the categories are ordered (such as "buy", "hold", "sell"), we can use methods seen in the context of generalied low-rank models.

Notes

Metrics

In regression, we can use average prediction error (on the test set) to evaluate a particular prediction algorithm.

In classification, we need to capture false positives and false negatives, and we can use similar metrics (evaluated on the *test* set):



- *Precision* p: the number of correctly predicted positive results divided by the number of all positive results,

$$p = \frac{TP}{TP + FP}.$$

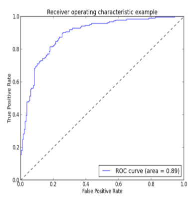
- *Recall* r: the number of correct positive results divided by the number of positive results that should have been returned,

$$r = \frac{TP}{TP + FN}.$$

Notes

Capturing both precision and recall

Finance Data Science
7. Classification
MFE 230P, Summer 2017
Binary Classification
Basics
SVM
Logistic regression
Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness
References



- *F1 score:* harmonic mean of p and r, attempting to capture both precision and recall in one score.
- *ROC curve:* the area under the curve.

Notes

Support Vector Machines
Separable data

The data is linearly separable if there exist a linear classification rule that makes no error on the training set.

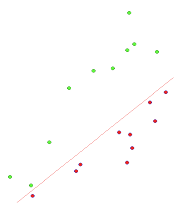
This is a set of linear inequalities constraints on (w, b):

$$y_i(w^T x_i + b) \geq 0, \quad i = 1, \ldots, n.$$

Strict separability corresponds the the same conditions, but with strict inequalities.

Notes

Geometry



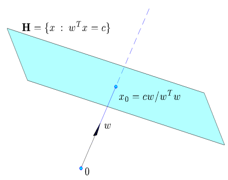
Geometrically: the hyperplane

$$\{x \ : \ w^T x + b = 0\}$$

perfectly separates the positive and negative data points.

Notes

Linear algebra flashback: hyperplanes



Geometrically, a hyperplane $\mathbf{H} = \left\{w \, : \, w^T x = c\right\}$ is a translation of the set of vectors orthogonal to w. The direction of the translation is determined by w, and the amount by $c/\|w\|_2$. Indeed, the projection of 0 onto $\mathbf{H}$ is $x_0 = cw/(w^T w)$.

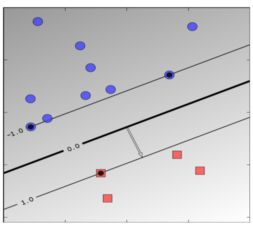
Notes

Geometry (cont'd)

Assuming strict separability, we can always rescale (w, b) and work with

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

Amounts to make sure that negative (resp. positive) class contained in half-space $w^T x + b \leq -1$ (resp. $w^T x + b \geq 1$).



The distance between the two "± 1" boundaries turns out the be equal to $2/\|w\|_2$.

Thus the "margin" $\|w\|_2$ is a measure of how well the hyperplane separates the data apart.

Finance Data Science
7. Classification
MFE 230P, Summer 2017
Binary Classification
Basics
SVM
Logistic regression
Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness
References

Notes

Non-separable data

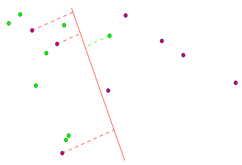
Separability constraints are homogeneous, so WLOG we can work with

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

If the above is infeasible, we try to minimize the "slacks"

$$\min_{w,b,s} \sum_{i=1}^{n} s_i \ : \ s \geq 0, \ \ y_i(w^T x_i + b) \geq 1 - s_i, \ \ i = 1, \ldots, n.$$

The above can be solved as a "linear programming" (LP) problem (in variables w, b, s).



Geometry of LP formulation.

Geometry of LP formulation: we minimize the sum of the distances from mis-classified points to the boundary.

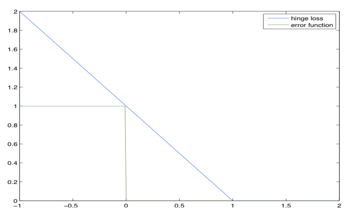
Notes

Hinge loss function

The previous LP can be interpreted as minimizing the hinge loss function

$$L(w, b) := \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b), 0).$$

This serves as an approximation to the number of errors made on the training set:



Notes

Handling class imbalance

In many applications, the number of positively labeled training points is much less than that of the negative class.

We can address the class imbalance issue via the modified loss:

$$L(w,b) := \frac{1}{m_+}\sum_{i\in\mathcal{I}_+}\max(1-y_i(w^Tx_i+b),0)+\frac{1}{m_-}\sum_{i\in\mathcal{I}_-}\max(1-y_i(w^Tx_i+b),0),$$

where $\mathcal{I}_\pm$ is the set of positively or negatively labelled points, and $m_\pm$ the corresponding number.

Notes

Regularization

The solution might not be unique, so we add a regularization term $\|w\|_2^2$:

$$\min_{w,b}\ \frac{1}{2}\|w\|_2^2 + C\cdot\sum_{i=1}^m \max(1-y_i(w^Tx_i+b),0)$$

where $C>0$ allows to trade-off the accuracy on the training set and the prediction error (more on why later). This makes the solution unique.

The above model is called the *Support Vector Machine* . It is a quadratic program (QP). It can be reliably solved using special fast algorithms that exploit its structure.

If C is large, and data is separable, reduces to the maximal-margin problem

$$\min_{w,b}\frac{1}{2}\|w\|_2^2\ :\ y_i(w^Tx_i+b)\geq 1,\quad i=1,\ldots,n.$$

Notes

Logistic regression

Logistic model

We model the probability of a label Y to be equal $y \in \{-1, 1\}$, given a data point $x \in \mathbf{R}^n$, as:

$$P(Y = 1 \mid x) = 1 - P(Y = -1 \mid x) = \frac{1}{1 + \exp(-(w^T x + b))}.$$

This amounts to modeling the *log-odds ratio* as a linear function of X:

$$\log \frac{P(Y = 1 \mid x)}{P(Y = -1 \mid x)} = w^T x + b.$$

- The decision boundary (the set of points x such that $P(Y = 1 \mid x) = P(Y = -1 \mid x)$) is the hyperplane with equation $w^T x + b = 0$.
- The region $P(Y = 1 \mid x) \geq P(Y = -1 \mid x)$ (*i.e.*, $w^T x + b \geq 0$) corresponds to points with predicted label $\hat{y} = +1$.

Notes

Maximum-likelihood

The likelihood function is

$$l(w, b) = \prod_{i \,:\, y_i = +1} \frac{1}{1 + e^{-(w^T x_i + b)}} \prod_{i \,:\, y_i = -1} \frac{e^{-(w^T x_i + b)}}{1 + e^{-(w^T x_i + b)}}.$$

Now maximize the log-likelihood:

$$\max_{w,b} \; L(w, b) := -\sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i + b)})$$

- In practice, we may consider adding a regularization term

$$\max_{w,b} \; L(w, b) + \lambda \|w\|_2^2.$$

- Many packages exist for logistic regression, *e.g.* [4].

Notes

Generalized classification

We consider the problem

$$\min_w \mathcal{L}(X^T w + b\mathbf{1}, y) + \lambda p(w),$$

where

- $\mathcal{L}$ is a convex loss function that encodes the error between the observed value and the predicted value;
- (w, b) are the model parameters;
- p is a penalty on the regression parameters;
- $\lambda > 0$ is a penalty parameter.

When $\mathcal{L}(z, y) = \mathbf{1}^T(1 - yz)_+$, $p(w) = \|w\|_2^2$, we recover regularized SVM.

Notes

Playing with loss functions and penalties

Changing loss functions allos to cover these types of regression methods:

- SVMs
- Logistic regression
- Naïve Bayes classification

Typical penalties allow to

- l_1-norm: to enforce sparsity;
- l_2-norm (often, squared): to control statistical noise and improve prediction error;
- sum-block norms enable to enforce whole blocks of w to be zero.

Notes

Motivations

In some applications, we have access to a measure of uncertainty associated with each data point, and model this as $X \in \mathcal{X}$, with $\mathcal{X}$ a matrix set that describe the uncertainty around a given data set $\hat{X} \in \mathcal{X}$.

Robust model:

$$\min_{w,b} \max_{X \in \mathcal{X}} \mathcal{L}(X^T w + b\mathbf{1}, y).$$

Notes

Example: interval model

Assume that each entry in the data matrix is only known to belong to a given interval:

$$X_{ij} \in [\hat{X}_{ij} - R_{ij}, \hat{X}_{ij} + R_{ij}],$$

with $\hat{X}_{ij}, R_{ij} > 0$ given, $1 \le i \le n$, $1 \le j \le m$.

This corresponds to the robust model

$$\min_{w,b} \max_{X \in \mathcal{X}} \mathcal{L}(X^T w + b\mathbf{1}, y),$$

with $\mathcal{X} = [\hat{X} - R, \hat{X} + R]$ an interval matrix (here $R = (R_{ij})$).

Notes

Explicit form

Key fact: for given $\hat{x} \in \mathbf{R}^n$, $\rho \in \mathbf{R}^n_+$:

$$\max_{x \,:\, |x-\hat{x}| \le r} w^T x = w^T \hat{x} + r^T |w|,$$

where $|z|$ denotes the vector of magnitudes of elements in vector z.

Notes

Example

Robust SVM

For the SVM (hinge loss) case, we obtain

$$\min_{w,b} \sum_{i=1}^m \max(1 - y_i(w^T x_i + b) + R_i^T |w|, 0),$$

where R_i stands for the i-th colum of R. This provides some form of l_1-regularization.

The above can be further approximated with the upper bound

$$\min_{w,b} \sum_{i=1}^m \max(1 - y_i(w^T x_i + b), 0) + \sigma^T |w|,$$

with $\sigma := \sum_i R_i$.

Notes

Ellipsoidal uncertainty

Another model involves a spherical (or more generally ellipsoidal) uncertainty, where each data point x_i is only known to belong to a sphere of center $\hat{x}_i$ and radius r_i. More generally:

$$x_i = \hat{x}_i + r_i D u_i$$

with $D = \mathbf{diag}(\sigma_1, \ldots, \sigma_n)$ is a positive-definite diagonal scaling matrix, and $r_i > 0$. (Intuition: up to a point-dependent scaling factor r_i, variances are the same across the data points.)

For the SVM (hinge loss) case, we obtain

$$\min_{w,b} \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b) + r_i \|D^w\|_2, 0),$$

This provides some form of l_2-regularization. Model can be further approximated by some form of standard l_2-norm regularized SVM:

$$\min_{w,b} \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b), 0) + \lambda \|Dw\|_2, \quad \lambda := \sum_i r_i.$$

This provides guidance on which scaled penalty to use, and also explains why normalizing data by variance may be beneficial. (Why?)

Notes

Robustness interpretation of SVM

Return to separable data in the SVM setup. The set of constraints

$$y_i(w^T x_i + b) \geq 0, \quad i = 1, \ldots, n,$$

has many possible solutions (w, b).

We will select a solution based on the idea of robustness (to changes in data points).

Notes

Maximally robust separating hyperplane

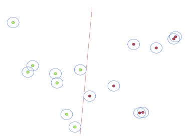
Spherical uncertainty model: assume that the data points are actually unknown, but bounded:

$$x_i \in \mathcal{S}_i := \{\hat{x}_i + u_i \ : \ \|u_i\|_2 \le \rho\},$$

where $\hat{x}_i$'s are known, $\rho > 0$ is a given measure of uncertainty, and u_i is unknown.

Robust counterpart: we now ask that the separating hyperplane separates the spheres (and not just the points):

$$\forall x_i \in \mathcal{S}_i \ : \ y_i(w^T x_i + b) \ge 0, \quad i = 1, \ldots, n.$$



For separable data we can try to separate spheres around the given points. We'll grow the spheres' radius until sphere separation becomes impossible.

Notes

Robust classification

Finance Data Science
7. Classification
MFE 230P, Summer 2017

Binary Classification
Basics
SVM
Logistic regression
Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness
References

We obtain the equivalent condition

$$y_i(w^T \hat{x}_i + b) \ge \rho \|w\|_2, \quad i = 1, \ldots, n.$$

Now we seek (w, b) which maximize ρ subject to the above.

By homogeneity we can always set $\rho\|w\|_2 = 1$, so that problem reduces to

$$\min_w \ \|w\|_2 \ : \ y_i(w^T \hat{x}_i + b) \ge 1, \quad i = 1, \ldots, n.$$

This is exactly the same problem as the SVM in separable case, a.k.a. the "maximum-margin classifier".

Notes

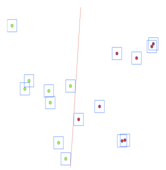
Separating boxes instead of spheres

We can use a box uncertainty model:

$$x_i \in \mathcal{B}_i := \{\hat{x}_i + u_i \;:\; \|u_i\|_\infty \le \rho\}.$$

This leads to

$$\min_w \; \|w\|_1 \;:\; y_i(w^T\hat{x}_i + b) \ge 1, \;\; i = 1, \ldots, n.$$



Classifiers found that way tend to be sparse. In 2D, the boundary line tends to be vertical or horizontal.

Notes

References

Finance Data Science
7. Classification

MFE 230P, Summer 2017

Binary Classification
Basics
SVM
Logistic regression

Regularization, sparsity, robustness
General model
Robustness
Sparsity and robustness

References

Sahely Bhadra, J Nath, Aharon Ben-Tal, and Chiranjib Bhattacharyya.
Interval data classification under partial information: A chance-constraint approach.
Advances in Knowledge Discovery and Data Mining, pages 208–219, 2009.

Chih-Chung Chang and Chih-Jen Lin.
LIBSVM: A library for SVM classification.

T. Hastie, R. Tibshirani, and J.H. Friedman.
The elements of statistical learning.
Springer, 2009.

David Madigan and David Lewis.
The BBR machine learning package, 2011.

Carolin Strobl, James Malley, and Gerhard Tutz.
An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.
Psychological methods, 14(4):323, 2009.

Notes