

Yanjun Qi / UVA CS 4501-01-6501-07

UVA CS 4501 - 001 / 6501 - 007

Introduction to Machine Learning and Data Mining

Lecture 8: k-CV on Regression & Classification with Support Vector Machine

Yanjun Qi / Jane

University of Virginia
Department of
Computer Science

9/18/14 1

Yanjun Qi / UVA CS 4501-01-6501-07

Where we are ? →

Five major sections of this course

- Regression (supervised)
- Classification (supervised)
- Unsupervised models
- Learning theory
- Graphical models

9/18/14 2

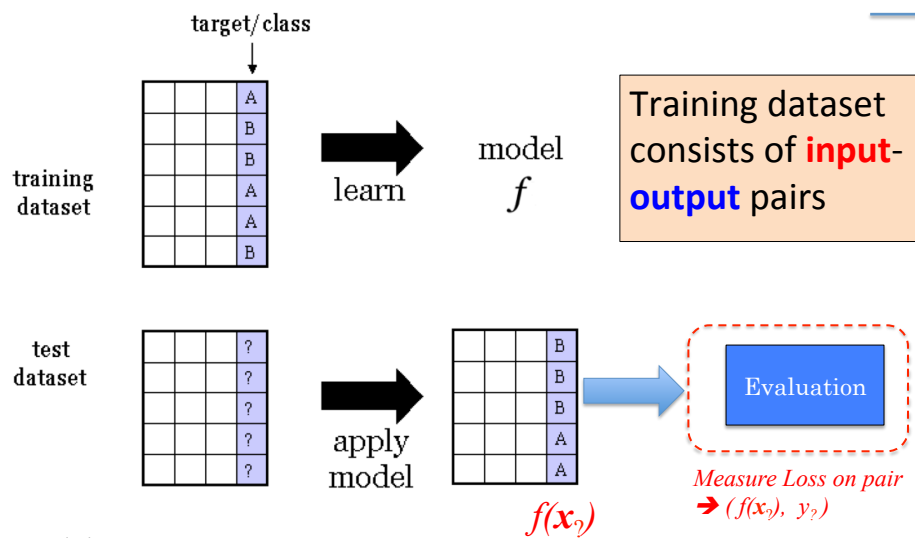
Today

- ❑ Review of basic pipeline
- ❑ Review of regression models
 - Linear regression (LR)
 - LR with non-linear basis functions
 - Locally weighted LR
 - LR with Regularizations
- ❑ Classification: Support Vector Machine (SVM)

9/18/14

3

SUPERVISED LEARNING

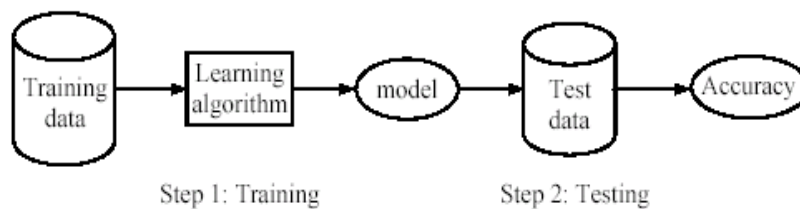


9/18/14

4

Evaluation Choice-I:

- ✓ **Training (Learning):** Learn a model using the training data
- ✓ **Testing:** Test the model using **unseen test data** to assess the model accuracy



$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

9/18/14

Evaluation Choice-II: e.g. 10 fold Cross Validation

- Divide data into 10 equal pieces
- 9 pieces as training set, the rest 1 as test set
- Collect the **scores** from the diagonal

model	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	train	train	train	train	train	train	train	train	train	test
2	train	train	train	train	train	train	train	train	test	train
3	train	train	train	train	train	train	train	test	train	train
4	train	train	train	train	train	train	test	train	train	train
5	train	train	train	train	train	test	train	train	train	train
6	train	train	train	train	test	train	train	train	train	train
7	train	train	train	test	train	train	train	train	train	train
8	train	train	test	train	train	train	train	train	train	train
9	train	test	train	train	train	train	train	train	train	train
10	test	train	train	train	train	train	train	train	train	train

9/18/14

HW2: Evaluation of Regression Models

training dataset \rightarrow

$$\mathbf{X}_{train} = \begin{bmatrix} -- & \mathbf{x}_1^T & -- \\ -- & \mathbf{x}_2^T & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n^T & -- \end{bmatrix} \quad \bar{\mathbf{y}}_{train} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

test dataset \rightarrow

$$\mathbf{X}_{test} = \begin{bmatrix} -- & \mathbf{x}_{n+1}^T & -- \\ -- & \mathbf{x}_{n+2}^T & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_{n+m}^T & -- \end{bmatrix} \quad \bar{\mathbf{y}}_{test} = \begin{bmatrix} y_{n+1} \\ y_{n+2} \\ \vdots \\ y_{n+m} \end{bmatrix}$$

9/18/14

7

HW2: Evaluation of Linear Regression Models

- Training Error:

$$J_{train}(\theta) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2$$

- Minimize $J_{train}(\theta) \rightarrow$ Normal Equation to get

$$\theta^* = \operatorname{argmin} J_{train}(\theta) = \left(\mathbf{X}_{train}^T \mathbf{X}_{train} \right)^{-1} \mathbf{X}_{train}^T \bar{\mathbf{y}}_{train}$$

9/18/14

8

HW2: Evaluation of Linear Regression Models

- Testing Error:

$$J_{test} = \frac{1}{2} \sum_{i=n+1}^{n+m} (\mathbf{x}_i^T \boldsymbol{\theta}^* - y_i)^2$$

HW2: Evaluation of Ridge Regression Models

- Training Error (regularized):

$$J_{train,\lambda}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\theta} - y_i)^2 + \lambda \sum_{j=1}^p \theta_j^2$$

- Minimize $J_{train,\lambda}(\boldsymbol{\theta}) \rightarrow$ (regularized) Normal Equation to get

$$\boldsymbol{\theta}_\lambda^* = \operatorname{argmin}_{\boldsymbol{\theta}} J_{train,\lambda}(\boldsymbol{\theta}) = \left(X_{train}^T X_{train} + \lambda I \right)^{-1} X_{train}^T \bar{y}_{train}$$

HW2: Evaluation of Ridge Regression Models

- Testing Error:

$$J_{test}(\lambda) = \frac{1}{2} \sum_{i=n+1}^{n+m} (\mathbf{x}_i^T \boldsymbol{\theta}_\lambda^* - y_i)^2$$

When varying tuning parameter λ ,

→ $J_{test}(\lambda)$ changes

→ So how to choose the best λ ?

Choose $\hat{\lambda}$ that generalizes well !

Today

- Review of basic pipeline
- Review of regression models
 - Linear regression (LR)
 - LR with non-linear basis functions
 - Locally weighted LR
 - LR with Regularizations
- Supervised Classification with Support Vector Machine (SVM)

Recap: Four Regression models

- Geometric view of four regression models we have learned
 - Linear regression (LR)
 - LR with non-linear basis functions
 - Locally weighted LR
 - LR with Regularizations

9/18/14

13

X_1	X_2	X_3	Y

A Dataset
for regression

$$f : X \rightarrow Y$$

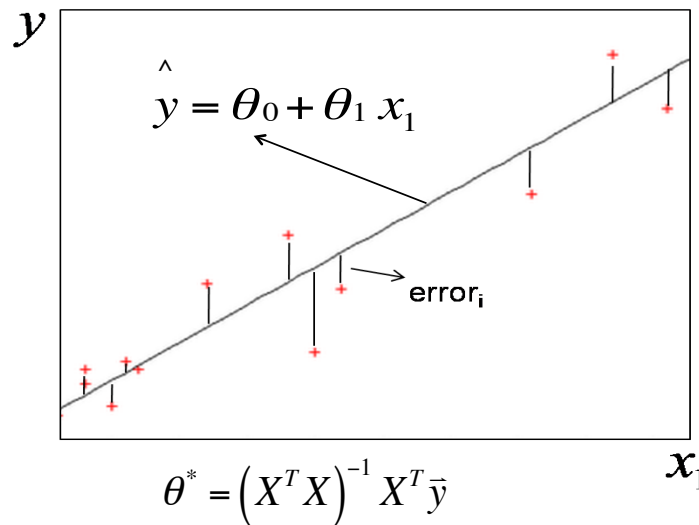
continuous
valued
variable

- **Data/points/instances/examples/samples/records:** [rows]
- **Features/attributes/dimensions/independent variables/covariates/predictors/regressors:** [columns, except the last]
- **Target/outcome/response/label/dependent variable:** special column to be predicted [last column]

9/18/14

14

1. Linear regression (LR)

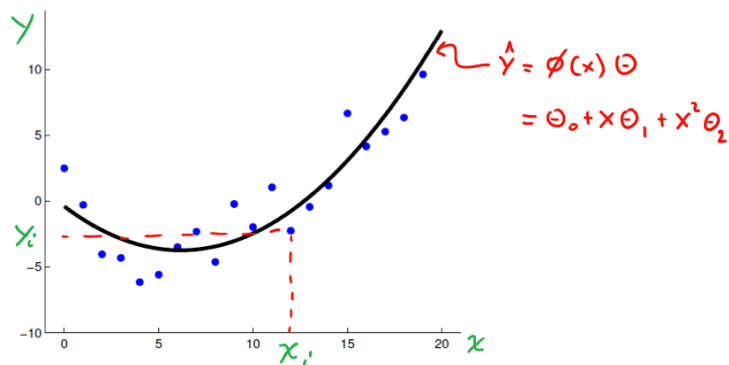


9/18/14

15

2. LR with nonlinear basis, e.g. Polynomial basis

For example, $\phi(x) = [1, x, x^2]$



$$y = \theta_0 + \sum_{j=1}^m \theta_j \phi_j(x) = \phi(x) \theta$$

9/18/14

16
Dr. Nando de Freitas's tutorial slide

2. LR with **nonlinear** basis, e.g Polynomial basis

$$\hat{y} = \theta_0 + \sum_{j=1}^m \theta_j \varphi_j(x) = \varphi(x)\theta$$

For example, $\varphi(x) := [1, x, x^2, x^3]$

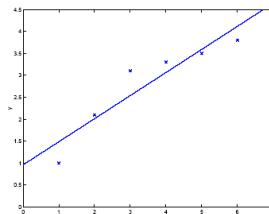
Another example, $\varphi(x) := [1, K_{\lambda=1}(x,1), K_{\lambda=1}(x,2), K_{\lambda=1}(x,4)]$

$$\theta^* = (\varphi^T \varphi)^{-1} \varphi^T \bar{y}$$

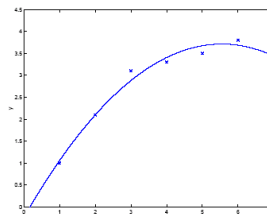
9/18/14

17

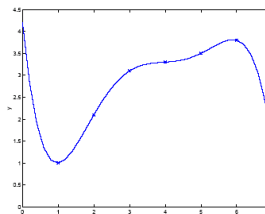
Which function *f to choose?*
Many possible choices , e.g.



$$y = \theta_0 + \theta_1 x$$



$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$



$$y = \sum_{j=0}^5 \theta_j x^j$$

Generalisation: learn function / hypothesis from **past data** in order to “explain”, “predict”, “model” or “control” **new data** examples

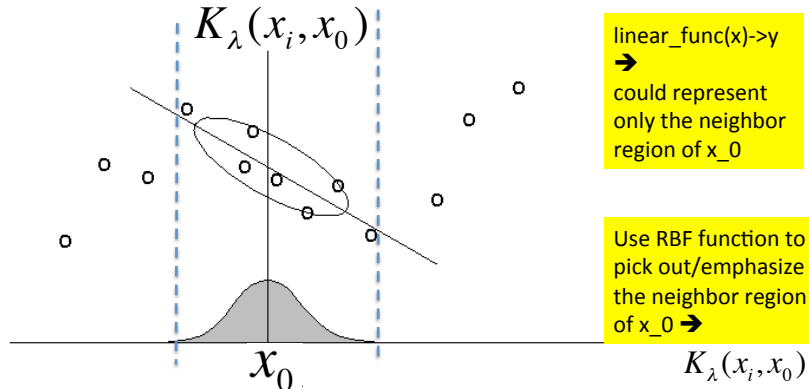
Choose *f* that generalizes well !

9/18/14

18

3. Locally Weighted LR

- aka locally weighted regression, locally linear regression, LOESS, ...

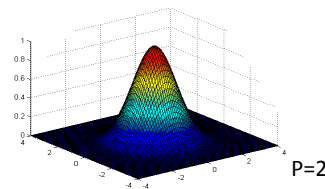
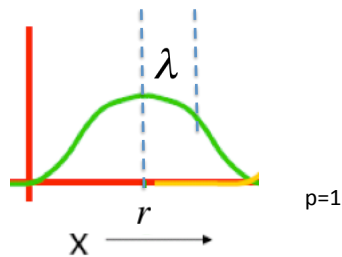


9/18/14 **Figure 2:** In locally weighted regression, points are weighted by proximity to the current x in question using a kernel. A regression is then computed using the weighted points.

RBF = radial-basis function: a function which depends only on the radial distance from a centre point

Gaussian RBF →
$$K_{\lambda}(x, r) = \exp\left(-\frac{(x-r)^2}{2\lambda^2}\right)$$

as distance from the centre r increases, the output of the RBF decreases



5. Regularized LR, e.g. Ridge regression

- The ridge estimator is solution from

$$J_{train,\lambda}(\theta) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2 + \lambda \sum_{j=1}^p \theta_j^2$$

to minimize $J(\theta)$, take derivative and set to zero

$$\theta^* = (X^T X + \lambda I)^{-1} X^T \bar{y}$$

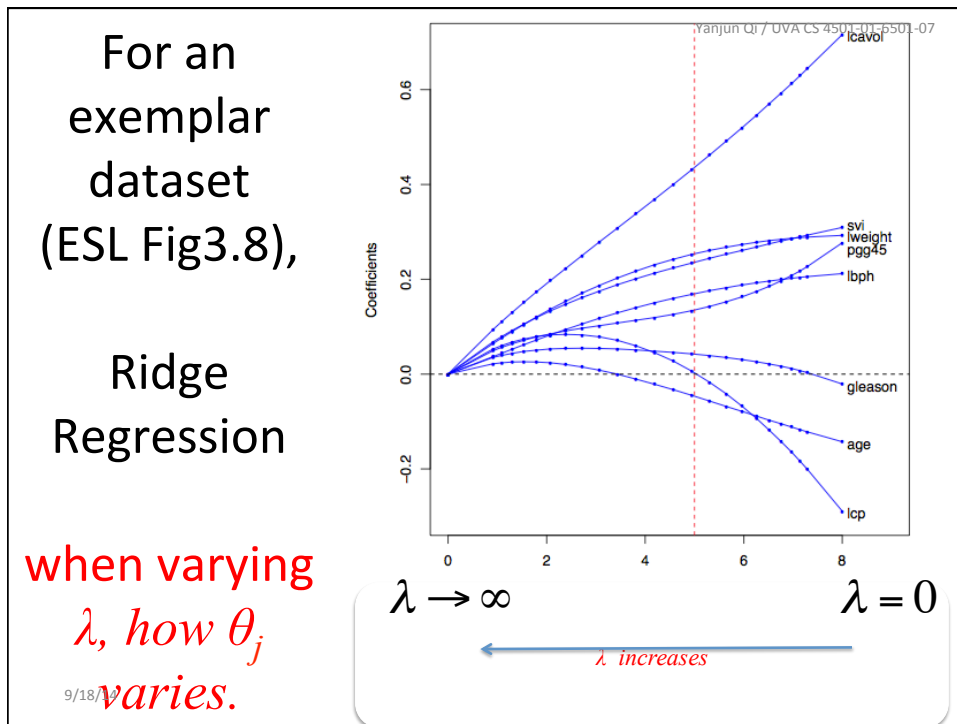
5. Regularized LR, e.g. Ridge regression

$$J_{train,\lambda}(\theta) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2 + \lambda \sum_{j=1}^p \theta_j^2$$

$$= \frac{1}{2} (X\theta - \bar{y})^T (X\theta - \bar{y}) + \lambda \theta^T \theta$$

$$= \frac{1}{2} (X\theta - \bar{y})^T (X\theta - \bar{y}) + \lambda \theta^T I \theta$$

$$= \frac{1}{2} (X\theta - \bar{y})^T (X\theta - \bar{y}) + \theta^T (\lambda I_{p \times p}) \theta$$



Yanjun Qi / UVA CS 4501-01-6501-07

Today

- Review of basic pipeline
- Review of regression models
 - Linear regression (LR)
 - LR with non-linear basis functions
 - Locally weighted LR
 - LR with Regularizations
- Supervised Classification**
- Support Vector Machine (SVM)

9/18/14 24

e.g. SUPERVISED LEARNING

Yanjun Qi / UVA CS 4501-01-6501-07

- Find function to map **input** space X to **output** space Y $f : X \longrightarrow Y$

- So that the **difference** between y and $f(x)$ of each example x is small.

e.g.

x	I believe that this book is not at all helpful since it does not explain thoroughly the material . it just provides the reader with tables and calculations that sometimes are not easily understood ...
----------	--

y	-1
----------	----

Output Y: {1 / Yes , -1 / No }
e.g. Is this a positive product review

Input X : e.g. a piece of English text

9/18/14

25

X_1	X_2	X_3	Y

X_1	X_2	X_3	Y

A Dataset for classification

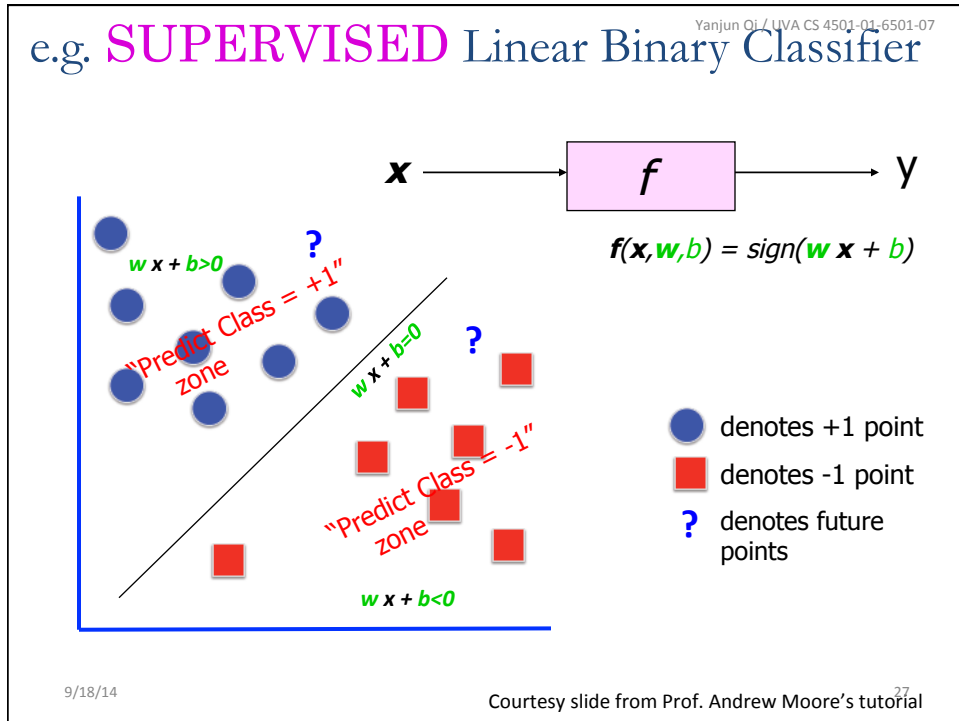
 $f : X \longrightarrow Y$

Output Class:
categorical
variable

- Data/points/instances/examples/samples/records:** [rows]
- Features/attributes/dimensions/independent variables/covariates/predictors/regressors:** [columns, except the last]
- Target/outcome/response/label/dependent variable:** special column to be predicted [last column]

9/18/14

26



Yanjun Qi / UVA CS 4501-01-6501-07

Classification: Application 1

- Direct Marketing
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before.
 - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

9/18/14 28

Classification: Application 2

- Fraud Detection
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

From [Berry & Linoff] Data Mining Techniques, 1997

9/18/14

29

Classification: Application 3

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where she/he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

9/18/14

30

Classification: Application 4

- Sky Survey Cataloging
 - Goal: To predict **class (star or galaxy) of sky objects**, especially visually faint ones, based on **the telescopic survey images** (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
 - Approach:
 - Segment the image.
 - Measure image attributes (features) - 40 of them per object.
 - Model the class based on these features.
 - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

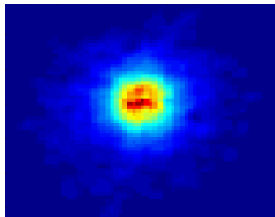
9/18/14

31

Classifying Galaxies

Courtesy: <http://aps.umn.edu>

Early



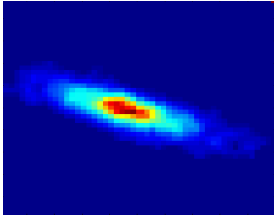
Class:

- Stages of Formation

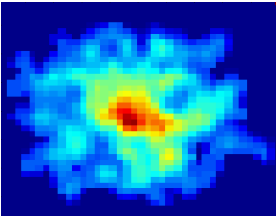
Attributes:

- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

From [Berry & Linoff] Data Mining Techniques, 1997

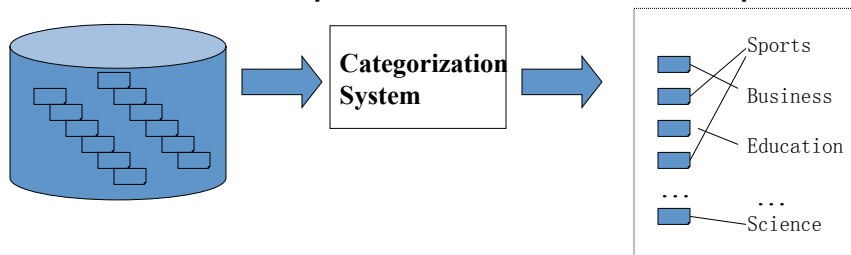
9/18/14

32

Yanjun Qi / UVA CS 4501-01-6501-07

Classification: Application 5 - Text Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new text documents
- A standard supervised classification problem



$\hat{y} = f(x)$

9/18/14 33

Yanjun Qi / UVA CS 4501-01-6501-07

Classification: Application 6 - Label Images into predefined classes

Motorbikes	Airplanes	Faces	Cars (Side)	Cars (Rear)	Spotted Cats	Background

9/18/14 34

Types of classifiers

- We can divide the large variety of classification approaches into roughly three major types
 1. Discriminative
 - directly estimate a decision rule/boundary
 - e.g., support vector machine, decision tree
 2. Generative:
 - build a generative statistical model
 - e.g., Bayesian networks
 3. Instance based classifiers
 - Use observation directly (no models)
 - e.g. K nearest neighbors

A study comparing Classifiers

An Empirical Comparison of Supervised Learning Algorithms

Rich Caruana
Alexandru Niculescu-Mizil

Department of Computer Science, Cornell University, Ithaca, NY 14853 USA

CARUANA@CS.CORNELL.EDU
ALEXN@CS.CORNELL.EDU

Abstract

A number of supervised learning methods have been introduced in the last decade. Unfortunately, the last comprehensive empirical evaluation of supervised learning was the Statlog Project in the early 90's. We present a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. We also examine the effect that calibrating the models via Platt Scaling and Isotonic Regression has on their performance. An important aspect of our study is the use of a variety of performance metrics to

This paper presents results of a large-scale empirical comparison of ten supervised learning algorithms using eight performance criteria. We evaluate the performance of SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps on eleven binary classification problems using a variety of performance metrics: accuracy, F-score, Lift, ROC Area, average precision, precision/recall break-even point, squared error, and cross-entropy. For each algorithm we examine common variations, and thoroughly explore the space of parameters. For example, we compare ten decision tree styles, neural nets of many sizes, SVMs with many kernels, etc.

Because some of the performance metrics we examine

A study comparing Classifiers

→ 11 binary classification problems

Yanjun Qi / UVA CS 4501-01-6501-07

PROBLEM	#ATTR	TRAIN SIZE	TEST SIZE	%POZ
ADULT	14/104	5000	35222	25%
BACT	11/170	5000	34262	69%
COD	15/60	5000	14000	50%
CALHOUS	9	5000	14640	52%
COV_TYPE	54	5000	25000	36%
HS	200	5000	4366	24%
LETTER.P1	16	5000	14000	3%
LETTER.P2	16	5000	14000	53%
MEDIS	63	5000	8199	11%
MG	124	5000	12807	17%
SLAC	59	5000	25000	50%

9/18/14

37

A study comparing Classifiers

→ 11 binary classification problems / 8 metrics

Yanjun Qi / UVA CS 4501-01-6501-07

Table 2. Normalized scores for each learning algorithm by metric (average over eleven problems)

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	.939	.963	.938	.929*	.880	.896	.896	.917
RF	PLT	.872*	.805	.934*	.957	.931	.930	.851	.858	.892	.898
BAG-DT	-	.846	.781	.938*	.962*	.937*	.918	.845	.872	.887*	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	-	.872	.790	.934*	.957	.931	.930	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	.861	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	.843*	.933*	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	-	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	.836*	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	-	.834*	.816	.939	.963	.938	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	-	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	-	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	-	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774

9/18/14

38

Today

- ❑ Review of basic pipeline
- ❑ Review of regression models
 - Linear regression (LR)
 - LR with non-linear basis functions
 - Locally weighted LR
 - LR with Regularizations
- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)

9/18/14

39

X_1	X_2	X_3	Y

A Dataset
for **binary**
classification

$$f : X \rightarrow Y$$

Output as
Binary Class:
only two
possibilities

- **Data/points/instances/examples/samples/records:** [rows]
- **Features/attributes/dimensions/independent variables/covariates/predictors/regressors:** [columns, except the last]
- **Target/outcome/response/label/dependent variable:** special column to be predicted [last column]

9/18/14

40

Yanjun Qi / UVA CS 4501-01-6501-07

Linear Classifiers

x

f

y_{est}

- denotes +1
- denotes -1

How would you classify this data?

9/18/14
41

Yanjun Qi / UVA CS 4501-01-6501-07

Linear Classifiers

x

f

y_{est}

- denotes +1
- denotes -1

How would you classify this data?

9/18/14
42

Yanjun Qi / UVA CS 4501-01-6501-07

Linear Classifiers

x → f → y^{est}

- denotes +1
- denotes -1

How would you classify this data?

9/18/14 43

Yanjun Qi / UVA CS 4501-01-6501-07

Linear Classifiers

x → f → y^{est}

- denotes +1
- denotes -1

How would you classify this data?

9/18/14 44

Yanjun Qi / UVA CS 4501-01-6501-07

Linear Classifiers

x

y^{est}

- denotes +1
- denotes -1

Any of these would be fine..

..but which is best?

9/18/14
45

Yanjun Qi / UVA CS 4501-01-6501-07

Classifier Margin

x

y^{est}

- denotes +1
- denotes -1

Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

9/18/14
46

Yanjun Qi / UVA CS 4501-01-6501-07

Maximum Margin

x

f

y^{est}

• denotes +1
 ○ denotes -1

The **maximum margin linear classifier** is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Linear SVM

9/18/14 47

Yanjun Qi / UVA CS 4501-01-6501-07

Maximum Margin

x

f

y^{est}

• denotes +1
 ○ denotes -1

Support Vectors are those datapoints that the margin pushes up against

The **maximum margin linear classifier** is the linear classifier with the, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Linear SVM

9/18/14 48

References

- Big thanks to Prof. Ziv Bar-Joseph @ CMU for allowing me to reuse some of his slides
- Elements of Statistical Learning, by Hastie, Tibshirani and Friedman
- Prof. Andrew Moore @ CMU's slides
- UMN Data Mining Course Slides