2014 2014 12.8 2014 7630.5

[1] 2004 [2] 2005 [3] K-
NN 2006 [4] 2009 [5] K-NN K-
NN 550 200 =12 K-NN 2010 [6] “ ” 2011 13 SVM
NN C4.5 PART SVM 97.64% 83.1% C4.5 PART 97.5%

LDA

d_i w_j $r_{ij} = 0$ $r_{ij} = 1$
LDA Latent Dirichlet Allocation “ ”

[1] (TF)[]
CHI χ^2 χ^2 χ^2 i j χ^2

$$\chi_{ij}^2 = \frac{n \times (n_{11} \times n_{22} - n_{12} \times n_{21})^2}{(n_{11} + n_{12}) \times (n_{21} + n_{22}) \times (n_{11} + n_{21}) \times (n_{12} + n_{22})}$$

$n_{11}n_{12}n_{21}\ n_{22}$ i j i j i j i j n χ^2

Word2vec Google 2013 K Word2vec Mikolov Word2vec

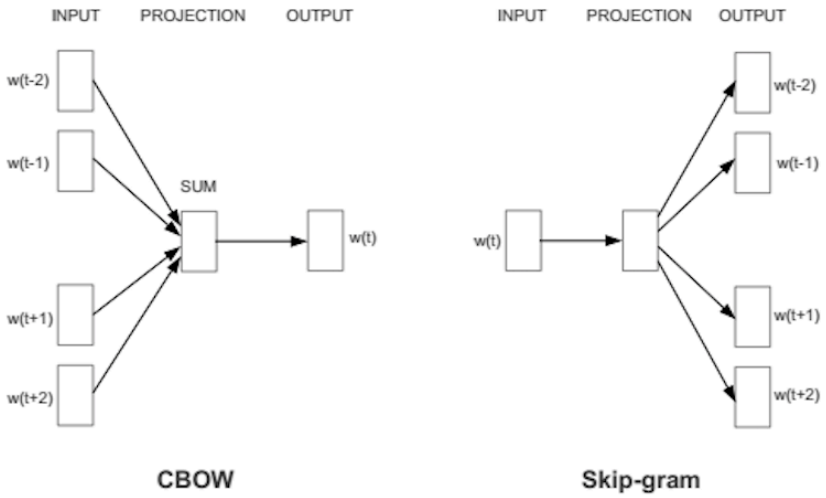


Figure 1

,CBOW Skip_gram CBOW Skip_gram Word2vec

RFC [10] $h(X, \Theta_k), k = 1, \ldots$ Θ_k X RFC boostrap RF k $h_1(X), h_2(X), \ldots, h_k(X)$:

$$H(x) = \arg\max_Y \sum_{i=1}^{k} I(h_i(x) = Y)$$

$H(x)$ h_i Y $I(*)$

RF RF

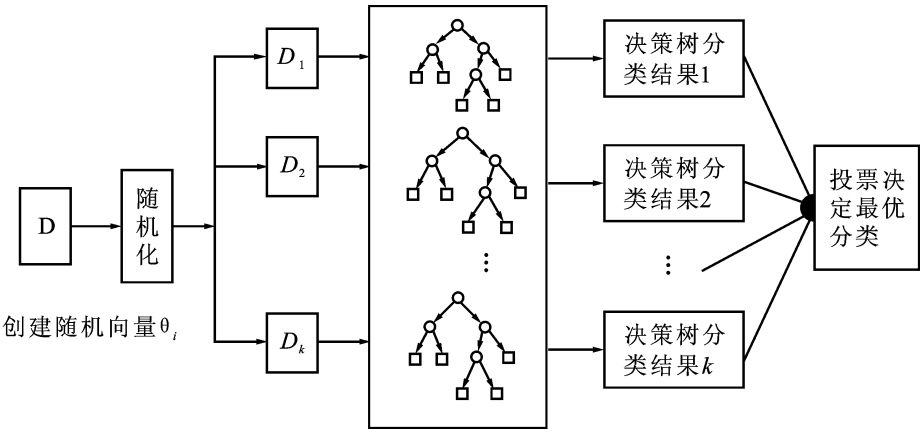
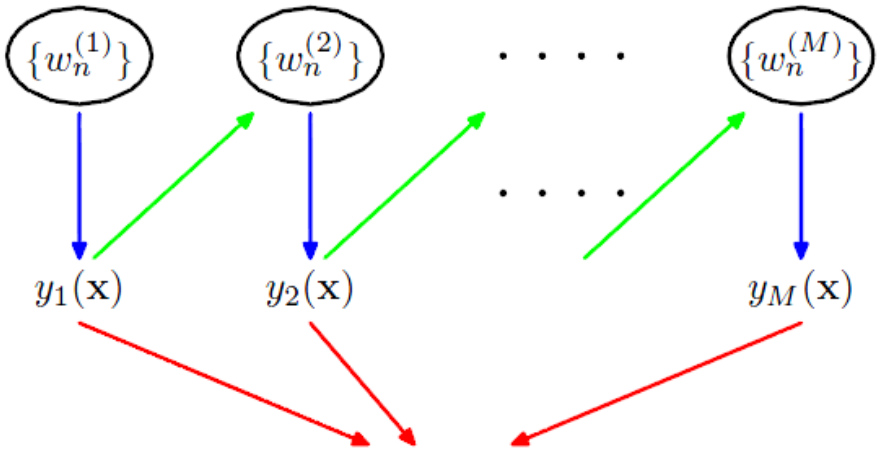


Figure 2

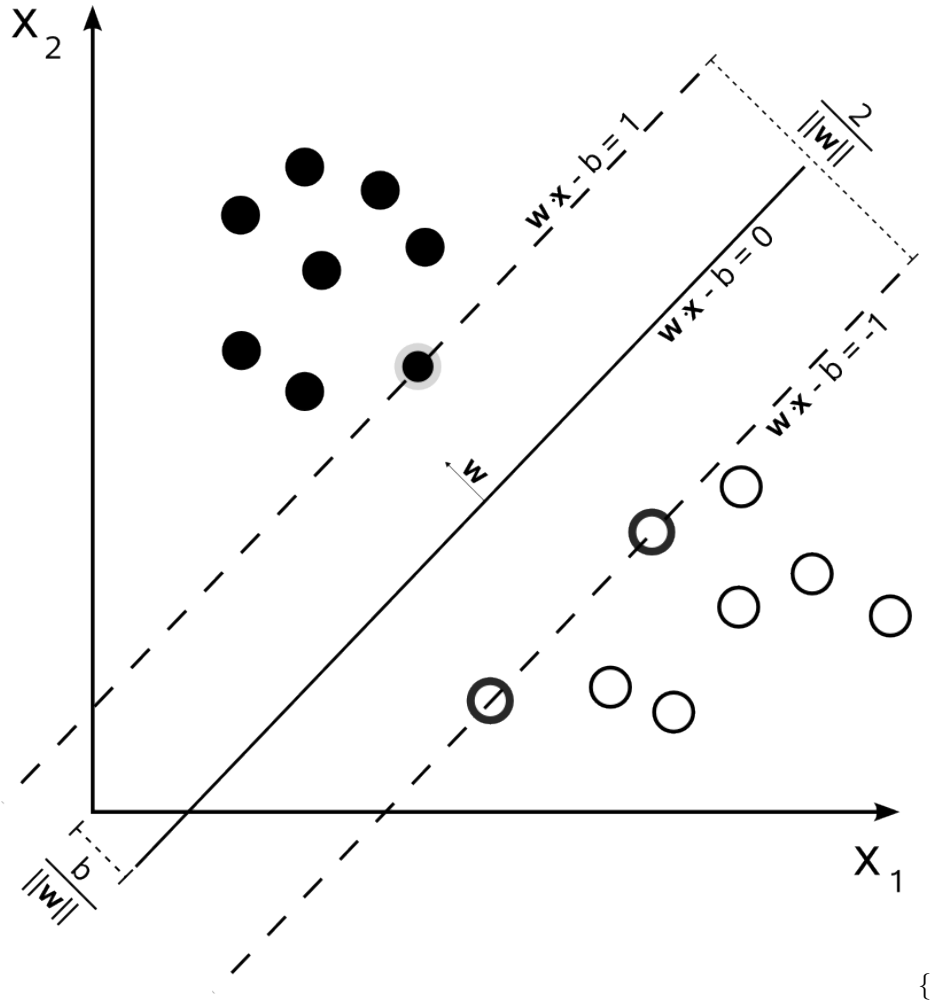
Gradient Boosting

Boosting Valiant (Probably Approximately Correct, PAC) Kearns Valiant[11] Boosting M



$$Y_M(\mathbf{x}) = \text{sign}\left(\sum_m^M \alpha_m y_m(\mathbf{x})\right)$$

Gradient Boosting Boosting (Loss Function) Gradient Boosting - Boosting



{#fig:svm,width:300px}

$(x_i, y_i), i = 1, 2, \ldots, l, x \in R^n, y \in 1, -1 \quad (\omega \cdot x_i) + b =$
0 ω n b $y_i(w \cdot x_i + b) \geq 1 \quad i = 1, 2, \ldots, n$
$2/|\omega|$

$$\min \Phi(\omega) = \frac{1}{2}|\omega|^2 = \frac{1}{2}\omega^T\omega$$

$L(\omega, b, a) = \frac{1}{2}|\omega| - a(y((\omega \cdot x) + b) - 1) \quad a_i > 0$ ω b

0

$$maxQ(a) = \sum_{j=1}^{l} a_j - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j (a_i \cdot x_j)$$
$$s.t. \quad \sum_{j=1}^{l} a_j y_j = 0 \qquad j = 1, 2, \ldots, l, a_j \geq 0, j = 1, 2, \ldots, l$$

$a^* = (a_1)$

$$\omega^* = \sum_{j=1}^{l} a_j^* y_j x_j$$

$$b^* = y_i - \sum_{j=1}^{l} y_j a_j^* (x_j \cdot x_i)$$

$j \in j | a_j^* > 0$ $(\omega^* \cdot x) + b^*$

$$f(x) = sign\{(\omega^* \cdot x) + b^*\} =$$
$$sign\{(\sum_{j=1}^{l} a_j^* y_j (x_j \cdot x_i)) + b^*\}, x \in R^n$$

x R^n H Φ

$$x \to \Phi(x) = (\Phi_1(x), \Phi_2(x), \ldots, \Phi_l(x))^T$$

$\Phi(x)$ x

$$f(x) = sign\{(\omega^* \cdot \Phi(x)) + b^*\} =$$
$$sign\{(\sum_{j=1}^{l} a_j^* y_j (\Phi(x_j) \cdot \Phi(x_i)) + b^*\}, x \in R^n$$

Φ , $< x_i, x_j >$

1. $K(x_i, K_j) = (x_1^T x_1)^d$
2. Gauss $K(x_i, K_j) = \exp(-q|x_1 - x_2|^2)$
3. B- Fourier

Logistic

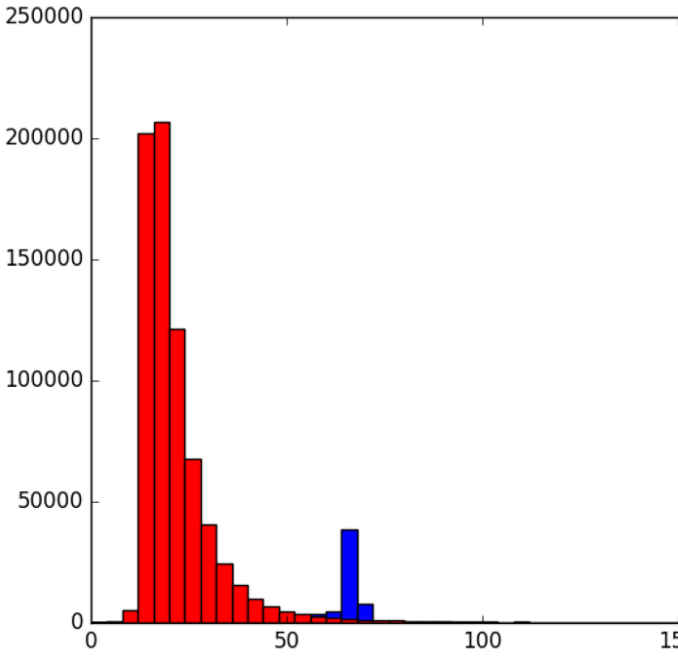
logistic generalized linear model $w'x$ + b w b $w'x+b$ $y = w'x+b$ logistic L $w'x+b$ p $p = L(w'x + b)$, p 1-p L logistic logistic L

Logistic

$$P(Y = 1|x) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)} P(Y = 0|x) = \frac{1}{1 + \exp(w \cdot x + b)}$$

TP(true positive)	FN(false negative)
FP(false positive)	TN(true negative)

1. Accuracy = $= \frac{TP+TN}{TP+FN+FP+TN}$
2. (Precision)= $= \frac{TP}{TP+FP}$
3. (Recall)= $= \frac{TP}{TP+FP}$
4. F1=$2 \times \frac{Precision \times Recall}{Precision+Recall}$



50 50 1/5

	precision	recall	f1-score	support
0.0	0.996999	0.999646	0.998320	143891
1.0	0.996757	0.973121	0.984797	16109
avg / total	0.996974	0.996975	0.996959	160000

{#fig:cjw_2,width:300px}

+

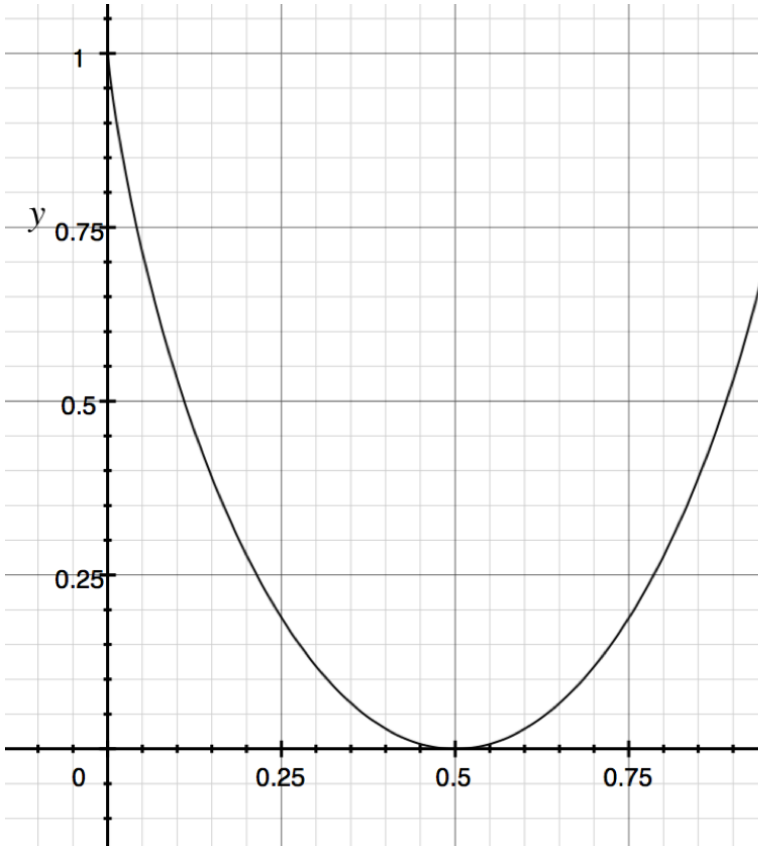
•

“ - ”

34

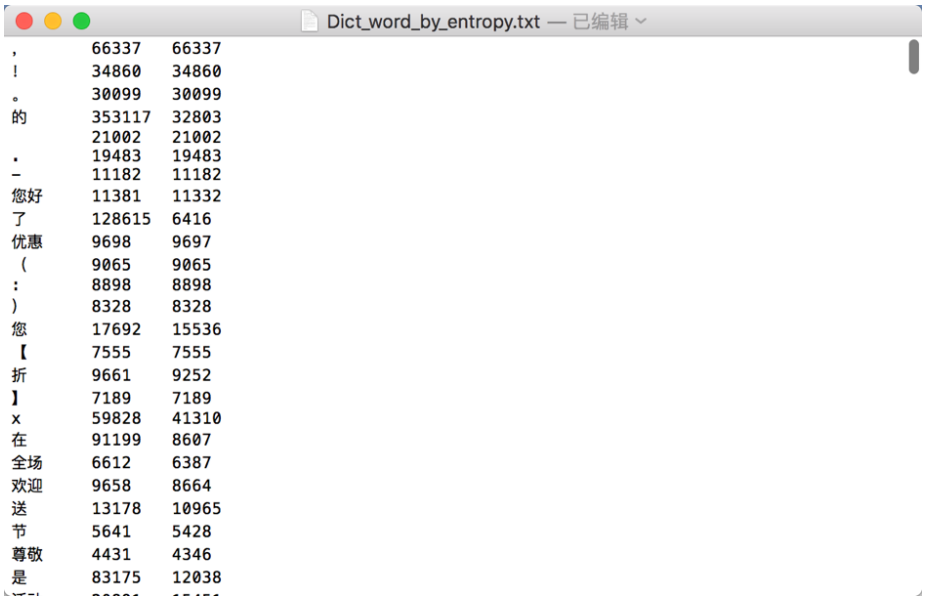
21

“ ” “ ” “0” “1”



$1+plog(p)+(1-p)log(1-p)$ f

$p > 0.5$ “1” “0” “$t = /1$” $t * f$



6.6 8 “ ”

-

-

= 1001 ntree = 10

Precision	Recall	F1
0.998715	0.999423	0.999069
0.994815	0.998516	0.991655
0.998323	0.998325	0.998323

= 1001 ntree = 100

Precision	Recall	F1
0.998826	0.999472	0.999149
0.995255	0.989509	0.992374
0.998467	0.998469	0.998467

= 101 ntree = 10

	Precision	Recall	F1
	0.996964	0.999666	0.998314
	0.996962	0.972810	0.984730
	0.996962	0.996962	0.996946

= 11 ntree = 10

	Precision	Recall	F1
	0.997061	0.999527	0.998292
	0.995683	0.973679	0.984558
	0.996922	0.996925	0.996910

“ ”

χ^2 +Gradient Boosting

- χ^2 χ^2 term χ^2 χ^2 5000
- Gradient Boosting xgboost xgboost eXtreme Gradient Boosting Gradient Boosting Machine c++ xgboost CPU
- χ^2 5000 xgboost 5-fold

	F1
1	0.987383925712
2	0.987745868158
3	0.986261345269
4	0.987070464388
5	0.987418438388
avg	0.987176008383

Parameter	Setting
objective	binary logistic
max_depth	50
num_boost_round	50
learning_rates	0.5

Word2vec+

•

Word2vec Word2vec word- word2vec 30-300

•

SVM sklearn C++ SVM SMO

•

sklearn svm 5 64 16 5 :

	precision	recall	F1
0	1.00	1.00	1.00
1	1.00	0.98	0.99
avg/total	1.00	1.00	1.00

	precision	recall	F1
0	1.00	1.00	1.00
1	1.00	0.98	0.99
avg/total	1.00	1.00	1.00

	precision	recall	F1
0	1.00	1.00	1.00
1	1.00	0.98	0.99
avg/total	1.00	1.00	1.00

	precision	recall	F1
0	1.00	1.00	1.00
1	1.00	0.98	0.99
avg/total	1.00	1.00	1.00

	precision	recall	F1
0	1.00	1.00	1.00
1	1.00	0.98	0.99
avg/total	1.00	1.00	1.00

	precision	recall	F1
1	0.98360346	0.9966771	0.9900971
2	0.98315658	0.99696356	0.9900119
3	0.98281485	0.99656576	0.9896425
4	0.98261791	0.98261791	0.9893053
5	0.98293515	0.9973555	0.9900928
avg	0.9830256	0.9967293	0.98982995

svm(C++) C++ SVM 200 100 word2vec 50 :

precision	recall	F1
0.970000	0.928571	0.948833

+Logistic/SVM

•

DF (Document Frequency) TF-IDF term frequency–inverse document frequency (Information Gain, IG)

TF-IDF 32k 10
32k 10

• Logistic SVM

•

10 5-fold

Logistic

	F1
1	0.9915
2	0.9914
3	0.9908
4	0.9919
5	0.9910
avg	0.9913

SVM

	F1
1	0.9921
2	0.9917
3	0.9930

	F1
4	0.9915
5	0.9922
avg	0.9921

F1 0.987 SVM 1 F1

	F1
	0.997
+	0.998
χ^2 +Gradient Boosting	0.987
Word2vec+	0.990
+Logistic	0.991
+	0.992

http://nd-fe.zale.site χ^2 +Gradient Boosting

{#fig:pos,width=300px}

{#fig:neg,width=300px}

TF-IDF SVM
LR SVM LR

“ ” 0.998 F1

χ^2 (Chi-
Square) xgboost F1 98.8%

word2vec 100 SVM C++ SVM

[1] . [D][D]. , 2009.

[2] XIANG Y, CHOWDHURY M, ALI S. Filtering mobile spam by support vector machine[C]//CSITeA'04: Third International Conference on Computer Sciences, Software Engineering, Information Technology, E-Business and Applications. International Society for Computers; Their Applications (ISCA), 2004: 1–4.

[3] HEALY M, DELANY S J, ZAMOLOTSKIKH A. An assessment of case base reasoning for short text message classification[C]//Conference papers. 2004: 42.

[4] GÓMEZ HIDALGO J M, BRINGAS G C, SÁNZ E P, . Content based SMS spam filtering[C]//Proceedings of the 2006 ACM symposium on Document engineering. ACM, 2006: 107–114.

[5] LONGZHEN D, AN L, LONGJUN H. A new spam short message classification[C]//2009 First International Workshop on Education Technology and Computer Science. 2009.

[6] LIU W, WANG T. Index-based online text classification for sms spam filtering[J]. Journal of Computers, 2010, 5(6): 844–851.

[7] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//ICML. 1997, 97: 412–420.

[8] RONGLU L, JIANHUI W, XIAOYUN C, . Using maximum entropy model for Chinese text categorization [J][J]. Journal of Computer Research and Development, 2005, 1: 22–29.

[9] SCHÜTZE H, HULL D A, PEDERSEN J O. A comparison of classifiers and document representations for the routing problem[C]//Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1995: 229–237.

[10] . [J]. , 2013, 50(4): 1190–1197.

[11] KEARNS M, VALIANT L. Cryptographic limitations on learning Boolean formulae and finite automata[J]. Journal of the ACM (JACM), ACM, 1994, 41(1): 67–95.