1. Classification
   (a) K-Nearest Neighbor
   (b) Support Vector Machines
   (c) Adaboost
   (d) Iterative Dichotomiser 3
   (e) C4.5
   (f) Naive Bayes
   (g) Bagging
   (h) Random Forest
2. Neural Networks
   (a) Perceptron
   (b) Back-Propagation
   (c) Learning Vector Quantization
   (d) Self Organizing Map
3. Clustering Algorithms
   (a) Hierarchial Agglomerative Clustering
   (b) Hierarchial Division Clustering
4. Regression Algorithms
   (a) Lasso Regression
   (b) Logistic Regression
5. Deep Learning Algorithms
   (a) Deep Q-Learning
6. Other Methods
   (a) Gradient Descent
   (b) Gaussian Process

Algorithm 1 k-Nearest Neighbor [Tay et al., 2014] link:36

Input: X: training data, Y:Class labels of X, x: unknown sample
Output: Class label of unknown sample ClassifyX, Y, x $i = 1$ to m
1: Compute distance $d(X_i, x)$
2: Compute set I containing indices for the k smallest distances $d(X_i, x)$
3: Return majority label $\{Y_i$ where $i \in I\}$

Algorithm 2 Adaboost [Schapire, 2014]

Input:

Training data $\{(x_i, y_i)_{i=1}^N$ where $x_i \in \mathbb{R}^k$ and $y_i \in \{-1, 1\}\}$

Large number of classifiers denoted by $f_m(x) \in \{-1, 1\}$

0-1 loss function I defined as

$$I(f_m(x, y)) = \begin{cases} 0, & \text{if } f_m(x_i) = y_i \quad (1) \\ 1, & \text{if } f_m(x_i) \neq y_i \quad (2) \end{cases}$$

Output: The final classifier

$i = 1$ to N $i = 1$ to M

Fit weak classifier m to minimize the objective function:

$\epsilon_m = \frac{\sum_{i=1}^N w_i^m I(f_m(x_i)) \neq y_i}{x^2+2x+1}$

where $I(f_m(x_i) \neq y_i) = 1$ if $f_m(x_i) \neq y_i$ and 0 otherwise

$\alpha_m = \ln \frac{1-\epsilon_m}{\epsilon_m}$

all i

$w_i^{m+1} = w_i^{(m)} e^{\alpha_m I(f_m(x_i) \neq y_i)}$

Algorithm 3 Adaboost [Hertzmann et al., 2015]

Input:

Training data $\{(x_i, y_i)_{i=1}^N$ where $x_i \in \mathbb{R}^k$ and $y_i \in \{-1, 1\}\}$

Output: Weighted sum that represents the final output of the boosted classifier

Given Training data $\{(x_i, y_i)$ where $y_i \in \{-1, 1\}\}$

initialize D_1 = uniform distribution on training examples $t = 1$ to T

Train weak classifier h_t on D_t

choose $\alpha_t > 0$

compute new distribution D_{t+1}: all i

multiply $D_t(x)$ by

$$\begin{cases} e^{-\alpha_t}, & (< 1) \text{ if } y_i = h_t(x_i) \quad (3) \\ e^{\alpha_t}, & (> 1) \text{ if } y_i \neq h_t(x_i) \quad (4) \end{cases}$$

renormalize

output final classifier $H_{final(x)} = sign(\sum \alpha_t h_t(x))$

Algorithm 4 Random forest [Bernstein, 2016] Link:39

Input: S: training set, F:Features and number of trees in forest B

Output: Constructed tree RANDOMFORESTS, F

$H \leftarrow \emptyset$ $i \in 1, ...B$

$S^{(i)} \leftarrow$ A bootstrap sample from S

$h_i \leftarrow RANDOMIZEDTREELEARN(S^i, F)$

$H \leftarrow H \bigcup \{h_i\}$

return H

RANDOMIZEDTREELEARNS, F

At each node:

$f \leftarrow$ a very small subset of F

Split on best feature in f

return The learned tree

Algorithm 5 Iterative Dichotomiser 3 [., 2015] Link:40

Input: D : Training Data, X : Set of Input Attributes
Output: A decision tree ID3D, X

Let T be a new tree all instances in D have the same class c
Label $(T) = c$; Return T $X = \emptyset$ or no attribute has positive information gain
Label (T) = most common class in D; Return T
$X \leftarrow$ attribute with highest information gain
Label$(T) = X$ each value x of X
$D_x \leftarrow$ instances in D with $X = x$ D_x is empty
Let T_x be a new tree
Label(T_x) = most common class in D
T_x = ID3$(D_x, X - \{x\})$
Add a branch from T to T_x labeled by x
return T

Algorithm 6 Perceptron [Brownlee, 2015d] Link:65

Input: $ProblemSize, InputPatterns, iterations_max, learn_rate$
Output: $Weights$

$i = 1$ to $iterations_{max}$
$Pattern_i \leftarrow SelectInputPattern(InputPatterns)$
$Activation_i \leftarrow ActivateNetwork(Pattern_i, Weights)$
$Output_i \leftarrow TransferActivation(Activation_i)$
$UpdateWeights(Pattern_i, Output_i, learn_{rate})$
Return $Weights$

Algorithm 7 Back-propagation [Brownlee, 2015a]

Input: $ProblemSize, InputPatterns, iterations_{max}, learn_{rate}$
Output: $Network$

$Network \leftarrow ConstructNetworkLayers()$
$Network_weights \leftarrow InitializeWeights(Network, ProblemSize)$
$i = 1$ to $iterations_{max}$
$Pattern_i \leftarrow SelectInputPattern(InputPatterns)$
$Output_i \leftarrow ForwardPropagate(Pattern_i, Network)$
$BackwardPropagateError(Pattern_i, Output_i, Network)$
$UpdateWeights(Pattern_i, Output_i, Network, learn_{rate})$
Return $Network$

Algorithm 8 Back-propagation1

Input:

Training Set $x^{(1)}, y^{(1)}, , ..., (x^{(m)}, y^{(m)})$

Output:

Gradient of the cost function

$\Delta_{ij}^{(l)} = 0$(for all l,i,j) $i = 1$ to m
Set $a^{(1)} = x^{(i)}$
Perform forward propagation to compute $a^{(l)}$ for $l = 2, 3, ...L$
Using $y^(i)$, compute $\delta^(L) = a^(L) - y^(i)$
Compute $\delta^(L-1), \delta^(L-2), ...\delta^(2)$
$\Delta_{ij}^{(l)} := \Delta_{ij}^{(l)} + a_j^{(l)} \delta_i^{(l+1)}$
$D_{ij}^{(l)} := \frac{1}{m}\Delta_{ij} + \lambda\theta_{ij}^{(l)}$ if $j \neq 0$
$D_{ij}^{(l)} := \frac{1}{m}\Delta_{ij}$ if $j = 0$

Algorithm 9 Learning Vector Quantization [Brownlee, 2015c]

Input: $ProblemSize, InputPatterns, iterations_{max}, CodebookVectors_{num}, learn_{rate}$
Output: $CodebookVectors$

$CodebookVectors \leftarrow InitializeCodebookVectors(CodebookVectors_{num}, ProblemSize)$
$i = 1$ to $iterations_{max}$
$Pattern_i \leftarrow SelectInputPattern(InputPatterns)$
$Bmu_i \leftarrow SelectBestMatchingUnit(Pattern_i, CodebookVectors)$ $Bmu_i^{attribute} \in Bmu_i$ $Bmu_i^{class} \equiv Pattern_i^{class}$
$Bmu_i^{attribute} \leftarrow Bmu_i^{attribute} + learn_{rate} \times (Pattern_i^{attribute} - Bmu_i^{attribute})$
$Bmu_i^{attribute} \leftarrow Bmu_i^{attribute} - learn_{rate} \times (Pattern_i^{attribute} - Bmu_i^{attribute})$
Return $CodebookVectors$

Algorithm 10 Self Organizing Map [Brownlee, 2015b]

Input: $InputPatterns, iterations_{max}, learn_{rate}, Grid_width, Grid_height$
Output: $CodebookVectors$

$CodebookVectors \leftarrow InitializeCodebookVectors(Grid_{width}, Grid_{height}, InputPatterns)$
$i = 1$ to $iterations_{max}$
$Learn_{rate}^i \leftarrow CalculateLearningRate(i, learn_{rate}^{init})$
$neighborhood_{size}^i \leftarrow CalculateNeighborhoodSize(i, neighborhood_{init}^{size})$
$Pattern_i \leftarrow SelectInputPattern(InputPatterns)$
$Bmu_i \leftarrow SelectBestMatchingUnit(Pattern_i, CodebookVectors)$
$Neighborhood \leftarrow Bmu_i$
$Neighborhood \leftarrow SelectNeighbors(Bmu_i, CodebookVectors, neighborhood_{size}^i)$ $Vector_i \in Neighborhood$ $Vector_i^{attribute} \in Vector_i$
$Vector_i^{attribute} \leftarrow Vector_i^{attribute} + learn_{rate} \times (Pattern_i^{attribute} - Vector_i^{attribute})$
Return $CodebookVectors$

Algorithm 11 Hierarchial Agglomerative Algorithm [Stein, 2016a]

Input:
$\langle V, E, w \rangle$.Weighted graph
d_c.Distance measure for two clusters
Output: $\langle V_T, E_T \rangle$.Cluster hierarchy or dendogram

$C = \{\{v \mid v \in V\}\}$ {Initial Clustering}
$V_t = \{v_C \mid C \in C\}, E_T = \emptyset$ {Initial Dendogram} $|C| > 1$
$update\_distance\_matrix(C, G, d_c)$
$\{C, C'\} = \underset{\{C_i, C_j\} \in C : C_i \neq C_j}{argmin} d_c(C_i, C_j)$
$C = (C \backslash \{C, C'\}) \cup \{C \cup C'\}$ {Merging}
$V_T = V_T \cup \{v_{C,C'}\}, E_T = E_T \cup \{\{v_{C,C'}, v_C\}, \{v_{C,C'}, v_C\}\}$ {Dendogram}
Return T

Algorithm 12 Hierarchial Divisive Algorithm [Stein, 2016b]

Input:
$\langle V, E, w \rangle$.Weighted graph
d_c.Distance measure for two clusters
Output: $\langle V_T, E_T \rangle$.Cluster hierarchy or dendogram

$C = \{V\}$ {Initial Clustering}
$V_t = \{v_C \mid C \in C\}, E_T = \emptyset$ {Initial Dendogram} $\exists C_x : (C_x \in C \wedge |C| > 1)$
$update\_distance\_matrix(C, G, d_c)$
$\{C, C'\} = \underset{\{C_i, C_j\} : C_i \cup C_j = C_x \wedge \ C_i \cap C_j = \emptyset}{argmax} d_c(C_i, C_j)$
$C = (C \backslash \{C, C'\}) \cup \{C \cup C'\}$ {Merging}
$V_T = V_T \cup \{v_{C,C'}\}, E_T = E_T \cup \{\{v_{C,C'}, v_C\}, \{v_{C,C'}, v_C\}\}$ {Dendogram}
Return T

Algorithm 13 C4.5 [Dai and Ji, 2014]

Input:
T : Training dataset
S : Attributes
Output: decision tree $Tree$ C4.5T T is $NULL$
return failure
S is $NULL$
return $Tree$ as a single node with most frequent class label in T
$|S| = 1$
return $Tree$ as a single node S
set $Tree = \{\}$
$a \in S$
set $Info(a,T) = 0$ and $SplitInfo(a,T) = 0$
compute $Entropy(a)$ $v \in values(a,T)$
set $T_{a,v}$as the subset of T with attribute $a = v$
$Info(a,T) += \frac{|T_{a,v}|}{|T_a|} Entropy(a)$
$SplitInfo(a,T) += -\frac{|T_{a,v}|}{|T_a|} \log \frac{|T_{a,v}|}{|T_a|}$
$Gain(a,T) = Entropy(a) - Info(a,T)$
$GainRatio(a,T) = \frac{Gain(a,T)}{SplitInfo(a,T)}$
set $a_{best} = argmax\{GainRatio(a,T)\}$
a_{best}into $Tree$ $v \in values(a_{best},T)$ call $C4.5(T_{a,v})$
return $Tree$

Algorithm 14 Gradient Descent

Input:
f
starting value x_1
termination tolerances
Output: $x_{maxIters}$
$i = 1$ to $maxIters$
Compute the search direction $d_t = -\delta f(x_t)$ $|d_T| < \epsilon_g$
return "Converged to critical point", output x_t
Find α_t so that $f(x_t + \alpha_t d_t) < f(x_t)$
$|\alpha_t d_T| < \epsilon_x$
return "Converged in x", output x_t
Find α_t so that $f(x_t + \alpha_t d_t) < f(x_t)$
Let $x_{t+1} = x_t + \alpha_t d_t$
Return "Max number of iterations reached", output $x_{maxIters}$

Algorithm 15 Naive Bayes

Input:
C : A fixed set of classes
D : Documents
Output: Category(Class) of the Documents TrainMultinomialNBC, D
$V \leftarrow EXTRACTVOCABULARY(D)$
$N \leftarrow COUNTDOCS(D)$ each $c \in C$
$N_c \leftarrow COUNTDOCSINCLASS(D,c)$
$prior|c| \leftarrow N_c/N$
$text_c \leftarrow CONCATENATETEXTOFALLDOCSINCLASS(D,C)$ each $t \in V$
$condprob|t||c| \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'+1})}$
return $V, prior, condprob$
ApplyMultinomialNB$C, D, prior, condprob, d$
$W \leftarrow EXTRACTTOKENSFROMDOC(V,d)$
each $c \in C$
$score|c| \leftarrow \log\ prior|c|$ each $t \in W$
$score|c| += \log condprob|t||c|$
return $arg\ max_{c \in C} score|c|$

Algorithm 16 Lasso Regression

Input:

ipy : Inner product vector, $ipy_i =< y, X_{\cdot i} >$

ipx : Inner product matrix, $ipx_{ij} =< X_{\cdot i}, X_{\cdot j} >$

λ : Penalty parameter

N : Number of samples

Output: $beta$: Regression parameter vector FastLassoipy, ipx, λ, N

1: `stop_thr` {Threshold for stopping iteration}
2: $p \leftarrow length(ipy)$
3: $beta \leftarrow 0$ with length p
4: $gc \leftarrow 0$ with length p $difBeta_{max} \geq$ `stop_thr`
5: $difBeta_{max} \leftarrow 0$
$j = 1 \leftarrow p$
6: $z \leftarrow (ipy|j| - gc|j|)/N + beta|j|$
7: `beta_tmp` $\leftarrow max(0, z - \lambda) - max(0, -z - \lambda)$
8: $difBeta \leftarrow$ `beta_tmp` $- beta|j|$
9: $difabs \leftarrow abs(difBeta)$ $difabs > 0$
10: $beta|j| \leftarrow$ `beta_tmp`
11: $gc \leftarrow gc + ipx|j| \times difBeta$
12: $difBeta_{max} = max(difBeta_{max}, difabs)$

Algorithm 17 Bagging

Input:

B: the number of bags or base hypotheses

L: Base Learning Algorithm

Output: New Training Sets Bagging$examples, B, L$ $i = 1$ to B

1: $examples_i \leftarrow$ a bootstrap sample of $examples$
2: Compute set I containing indices for the k smallest distances $d(X_i, x)$
3: $h_i \leftarrow$ apply$L$to$examples_i$
4: Return $h_1, h_2, ...h_B$

Algorithm 18 Deep Q-Learning with Experience Replay [Mnih et al., 2013]

Input:

D: data set

Q: Action-Value Function

Output: New Training Sets

$i = 1$ to M

1: Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi = \phi(s_1)$ $i = 1$ to T
2: With probability ϵ select a random action a_t otherwise select $a_t = max_a Q * (\phi(s_t).a : \theta)$
3: Execute action a_t in emulator and observe reward r and image x_{t+1}
4: Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
5: Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D
6: Set $y_j =$

$$\begin{cases} r_j, & \text{for terminal } \phi_{j+1} \quad (5) \\ r_j + \gamma max_{a'} Q(\phi_{j+1,a';\theta}), & \text{for terminal } \phi_{j+1} \quad (6) \end{cases}$$

7: Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to the following equation
8:

$$\Delta_\theta L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(.);s'\sim\epsilon[(r+\gamma max_{a'} Q(s',a';\theta_{i-1}) - Q(s,a;\theta_i))\Delta_{\theta_i} Q(s,a;\theta_i)]}$$

Algorithm 19 PageRank

Input:
G: inlink file
$iteration$: Number of iteration
Output: PageRank PageRank$G, iteration$
1: $d \leftarrow 0.85$ {damping factor: 0.85}
2: $oh \leftarrow G$ {get outlink hash from G}
3: $ih \leftarrow G$ {get inlink hash from G}
4: $N \leftarrow G$ {get number of pages from G} all p in the graph
5: $opg[p] \leftarrow \frac{1}{N}$ $iteration > 0$
6: $dp \leftarrow 0$ all p that has no out-links
7: $dp \leftarrow dp + d * \frac{opg[p]}{N}$ all p in the graph
8: $npg[p] \leftarrow dp + \frac{[1-d]}{N}$ all ip in $ih[p]$
9: $npg[p] \leftarrow dp + \frac{d*opg[ip]}{oh[ip]}$
10: $opg \leftarrow npg$
11: $iteration \leftarrow iteration - 1$

Algorithm 20 DBSCAN link:42

Input:
D: Data
ϵ:Threshold distance
$MinPts$: Minimum number of points required to form a cluster
Output: Clustered Data DBSCAN$D, \epsilon, minPts$
1: $C = 0$ each point Pin dataset$D$ P is visited
2: continue next point
3: mark P as visited
4: $NeighborPts = regionQuery(P, \epsilon)$ $sizeof(NeighborPts) < MinPts$
5: mark P as NOISE
6: $C =$ next cluster
7: $expandCluster(P, NeighborPts, C, \epsilon, MinPts)$
expandCluster$P, NeighborPts, C, \epsilon, MinPts$
8: add P to Cluster C each point $P'inNeighborPts$ P' is not visited
9: mark P' as visited
10: $NeighborPts' = regionQuery(P', eps)$ $sizeof(NeighborPts) >= MinPts$
11: $NeighborPts = NeighborPts$ joined with $NeighborPts'$ P' is not yet member of any cluster
12: add P' to cluster C
regionQueryP, ϵ
13: return all points within $P's$ ϵ $neighborhood$

Algorithm 21 Logistic Regression

Input:
Training data of the form $\{(x_1, 1), (x_2, 0), ..\}$
x: unknown sample
Output: The output is a probability that the given input point belongs to a certain class
1: $0 \leftarrow \beta$
2: Compute y by setting its elements to

$$y = \begin{cases} 1, & \text{if } g_i = 1 \quad (7) \\ 0, & \text{if } g_i = 2 \quad (8) \end{cases}$$

i = 1,2,..N
3: Compute p by setting its elements to

$$p(x_i, \beta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

i = 1,2,..N
4: Compute the diagonal matrix W. The ith diagonal element is $p(x_i, \beta)(1 - p(x_i; \beta))$
5: $z \leftarrow X\beta + W^{-1}(y - p)$
6: $\beta \leftarrow (X^T W X)^{-1} X^T W z$
7: If the stopping criteria,stop;otherwise go back to step 3

Algorithm 22 Gaussian Process

Input:

$X = \begin{bmatrix} x_1^T \\ \dots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times D}$, m training inputs

$y = \begin{bmatrix} y_1^T \\ \dots \\ y_n^T \end{bmatrix} \in \mathbb{R}^n$

$k(.,.) : \mathbb{R}^{D \times D}$

x_*test input

σ^2noise level on the observations

$$[y(x) = f(x) + \epsilon, \epsilon \sim N(0, \sigma^2)]$$

Output:

f_*

$cov(f_*)$

1: $K \in \mathbb{R}^{n \times n}$Gram matrix.$K_{ij} = k(x_i, x_j)$

$$k(x_*) = k_* = k(X, x_*) = \begin{bmatrix} k(x_1, x_*) \\ \dots\dots\dots \\ k(x_n, x_*) \end{bmatrix} \in \mathbb{R}$$

2: $\alpha = (K + \sigma_2 \mathbb{I}_n)^{-1} y$

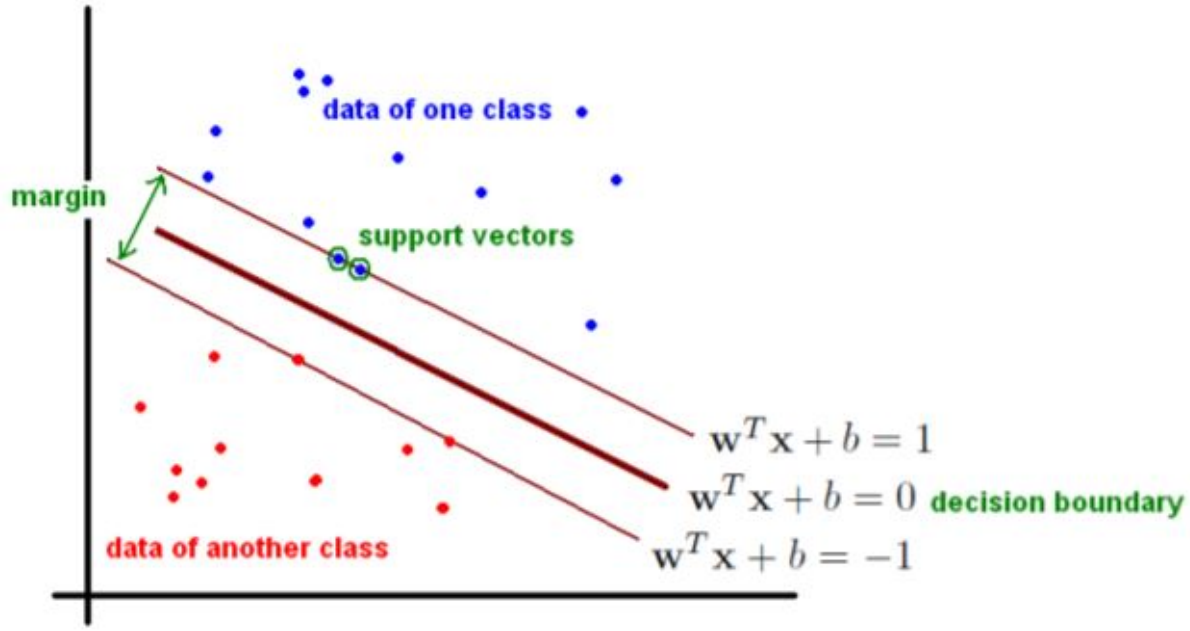
3: $f_* = k_*^T \alpha \in \mathbb{R}$

4: $cov(f_*) = k(x_*, x_*) - k_*^T [K + \sigma_2 \mathbb{I} n]^{-1} k_*$

Algorithm 23 Support Vector Machines

Input:

Set of N input-output pairs $\{x, y\}^{N_1}$ x: input vectors of the same dimension and y : set of output target labels $y_i = \{0, 1\}$



1 Simple case : linearly-separable data,binary classification

Goal:we want to find the hyperplane(i.e. decision boundary)linearly separating our classes. Our boundary will have the equation: $\mathbf{w^T x} + \mathbf{b} = 0$

Anything above the decision boundary should have label 1 i.e., $\mathbf{w^T x}_i + b > 0$ will have corresponding $y_i = 1$

Similarly, anything below the decision boundary should have label -1 i.e. $\mathbf{w^T x}_i + b < 0$ will have corresponding $y_i = -1$

The reason for this labeling scheme is that it lets us condense for the decision function to

$$f(x) = sign(\mathbf{w^T} + b)$$

since $f(x) = +1$ for all x above the boundary , and $f(x) = -1$ for all x below the boundary.

Thus, we can figure out if an instance has been classified properly by checking that $y(\mathbf{w^T} + b) \geq 1$ (which will be the case as long as either both $y, \mathbf{w^T} + b > 0$ or else $y, \mathbf{w^T} + b < 0$)

You'll notice that we will now have some space between our decision boundary and the nearest data points of either class. Thus,let's rescale the data such that anything on or above the boundary $\mathbf{w^T x} + b = 1$ is of one class(with label 1), and anything on or below the boundary $\mathbf{w^T x} + b = -1$ is of another class (with label -1)

What is the distance between these newly added boundaries?

First note that the two lines are parallel,and thus share their parameters w, b.Pick an arbitary point x_1 to lie on line $\mathbf{w^T x} + b = -1$. Then the closest point on line $\mathbf{w^T x} + b = 1$ is the point $\mathbf{x}_2 = \mathbf{x}_1 + \lambda\mathbf{w}$(since the closest point will always lie on the perpendicular;recall that the vector $\mathbf{w}$ is perpendicular to both lines). Using this formulation, $\lambda\mathbf{w}$ will be the line segment segment connecting $\mathbf{x}_1$ and $\mathbf{x}_2$, and thus, $\lambda \|\mathbf{w}\|$, the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is the shortest distance between the two lines/boundaries. Solving for λ : $\mathbf{w^T x}_2 + b = 1$ where $\mathbf{x}_2 = \mathbf{x}_1 + \lambda\mathbf{w}$
$\mathbf{w^T}(\mathbf{x}_1\lambda\mathbf{w}) + b = 1$

References

[., 2015] . (2015). Decision trees.

[Bernstein, 2016] Bernstein, M. (2016). Random forests.

[Brownlee, 2015a] Brownlee, J. (2015a). Back-propagation.

[Brownlee, 2015b] Brownlee, J. (2015b). Clever algorithms: Nature-inspired programming recipes.

[Brownlee, 2015c] Brownlee, J. (2015c). Learning vector quantization.

[Brownlee, 2015d] Brownlee, J. (2015d). Perceptron.

[Dai and Ji, 2014] Dai, W. and Ji, W. (2014). A mapreduce implementation of c4.5 decision tree algorithm.

[Hertzmann et al., 2015] Hertzmann, A., Fleet, D., and Brubaker, M. (2015). Adaboost.

[Mnih et al., 2013] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602.

[Schapire, 2014] Schapire, R. (2014). Machine learning algorithms for classification.

[Stein, 2016a] Stein, B. (2016a). Unit hierarchial cluster analysis.

[Stein, 2016b] Stein, B. (2016b). Unit hierarchial cluster analysis.

[Tay et al., 2014] Tay, B., Hyun, J., and Sejong, O. (2014). A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images.

Algorithm 23 Support Vector Machines (continued)

$\mathbf{w^T}(\mathbf{x}_1\lambda\mathbf{w}) + b = 1$
$\mathbf{w^T}\mathbf{x}_1 + b + \lambda\mathbf{w^T}\mathbf{w} = 1$ where $\mathbf{w^T}\mathbf{x}_1 + b = -1$
$-1 + \lambda\mathbf{w^T}\mathbf{w} = 1$
$\lambda\mathbf{w^T}\mathbf{w} = 2$
$\lambda = \frac{2}{\mathbf{w^T}\mathbf{w}} = \frac{2}{\|\mathbf{w}\|^2}$
And, so the distance $\lambda\|w\| is \frac{2}{\|\mathbf{w}\|^2}\|w\| = \frac{2}{\|w\|} = \frac{2}{\sqrt{\mathbf{w^T}\mathbf{w}}}$

It's intuitive that we would want to maximize the distance between the two boundaries demarcating the classes(Why?We want to be as sure that we are not making classification mistakes and thus we want our data points from the two classes to lie as far away from each other as possible).This distance is called the margin,so went to obtain the maximal margin.

Thus, we want to maximize $\frac{2}{\sqrt{\mathbf{w^T}\mathbf{w}}}$, which is equivalent to minimizing $\frac{\sqrt{\mathbf{w^T}\mathbf{w}}}{2}$ which is in turn equivalent to minimizing $\frac{\mathbf{w^T}\mathbf{w}}{2}$(since square root is a monotonic function)

This quadratic programming problem is expressed as :
$min_{w,b}\frac{\mathbf{w^T}\mathbf{w}}{2}$
subject to : $y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1(\forall$ data points $\mathbf{x}_i)$

2 Soft-margin extension

Consider the case that your data isn't linearly separable.For instance, maybe you aren't guaranteed that all your data points are correctly labelled, so you want to allow some data points of one class to appear on the other side of the boundary.

We can introduce $slackvariables$an $- \epsilon_i \geq 0$. Our quadratic programming problem becomes:
$min_{\mathbf{w},b,\epsilon}\frac{\mathbf{w^T}\mathbf{w}}{2} + C\sum_i \epsilon_i$
subject to : $y_i(\mathbf{w^T}\mathbf{x}_i + b) \geq 1 - \epsilon$

3 Nonlinear decision boundary

Mapping your data vectors, x_i,into a higher-dimension(even infinite) feature space may make them linearly separable in that space (whereas they may not be linearly separable in the original space). The formation of the quadratic programmatic problem is as above,but with all $\mathbf{x}_i$ replaced with $\phi(\mathbf{x}_i)$, where ϕprovides the higher-dimensional mapping.So we have the standard SVM formulation:
$min_{\mathbf{w},b,\epsilon}\frac{\mathbf{w}^T\mathbf{w}}{2}$
subject to : $y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \epsilon$and $\epsilon_i \geq 0(\forall$ data points $\mathbf{x}_i)$

4 Reformulating as a Lagrangian

We can introduce Lagrange multipliers to represent the condition:
$y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b))$ must be as close to 1 as possible. This condition is captured by: $max_{ai\geq 0}\alpha_i[1 - y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b))]$ This ensures that when $y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b)) \geq 1$, the expression above is maximal when $\alpha_i = 0$(since [1-$y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b)]$ ends up being negative). Otherwise, $y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b)) < 1$, so $[1 - y_i(\mathbf{w}^T\phi(\mathbf{x}_i + b))]$ is a positive value, and the expression is maximal when $a_i \to \infty$. This has the effect of penalizing any misclassified data points, while assigning 0 penalty to properly classified instances.

We thus have the following formulation:
$min_{w,b}[\frac{\mathbf{w}^T\mathbf{w}}{2} + \sum_i max_{\alpha\geq 0}\alpha_i[1 - \mathbf{w}^T\phi(\mathbf{x}_i + b)]]$
To allow for slack(soft-margin), preventing the α variables from going to ∞, we can impose constraints on the Lagrange multipliers to lie within: $0 \leq \alpha_i \leq C$. We can define the dual problem by interchanging the max and min as follows (i.e minimize after fixing alpha):
$max_{alpha\geq zero}[min_{w,b}J(\mathbf{w}, b; \alpha)]$where $= \frac{\mathbf{w}^T\mathbf{w}}{2} + \sum_i \alpha_i[1 - y_i(\mathbf{w}^T\phi(x_i) + b)]$

Algorithm 23 Support Vector Machines (continued)

Since, we're solving an optimization problem, we set $\frac{\partial J}{\partial \mathbf{w}} = 0$ and discover that the optimal setting of $\mathbf{w}$ is $\sum_i \alpha_i y_i \phi(x_i)$,while seeting $\frac{\partial J}{\partial \mathbf{b}} = 0$ yields the constraint $\sum_i \alpha_i y_i = 0$

Thus, after substituting and simplifying,we get:
$min_{w,b} J(\mathbf{w}, b, \alpha) = \sum_i \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ And thus our dual is:
$max_{\alpha \geq 0} [\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)]$
Subject to: $\sum_i \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$

5 Kernel trick

Because we're working in a higher-dimension space(and potentially even an infinite-dimensional space),calculating $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ may be intractable.However, it turns out there are special *kernel* functions that operate on the lower dimension vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ to produce a value equivalent to the dot-product of the higher dimensional vectors. For instance, consider the function $\phi: \mathbb{R}^3 \longmapsto \mathbb{R}^{10}, where$: $\phi(x) = (1, \sqrt{2}x^{(1)}, \sqrt{2}x^{(2)}, \sqrt{2}x^{(3)}, [\mathbf{x}^{(1)}]^2, [\mathbf{x}^{(2)}]^2, [\mathbf{x}^{(3)}]^2, \sqrt{2}x^{(1)(2)}, \sqrt{2}x^{(1)(3)}, \sqrt{2}x^{(2)(3)})$ Instead, we have the the kernel trick, which tells us that $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2 = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ for the given ϕ. Thus, we can simplify our calculations. Re-writing the dual in terms of the kernel yields: $max_{\alpha \geq 0} [\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j]$

6 Decision function

To classify a novel instance $\mathbf{x}$ once you've learned the optimal α_i parameters, all you have to do is calculate $f(\mathbf{x}) = sign(\mathbf{w}^T \mathbf{x} + b) = \sum_i \alpha_i y_i \phi(\mathbf{x}_i$ and using the kernel trick). Note that α_i is only non-zero for instances $\phi(\mathbf{x}_i)$ on or near the boundary-those are called the *supportvector* since they alone specify the decision boundary. We can toss out the other data points once training is complete. Thus, we only sum over the x_i which constitute the support vectors.