Data Mining algorithms

2017-2018 spring

03.07.2018

1. Maximal margin
2. Generalization

Plan

W1 Februar 7-9: Introduction, kNN, evaluation

W2 Februar 14-16: Evaluation, Decision Trees

W3 Februar 21-23: Linear separators, iPython, VC theorem

W4 Februar 28-march 2: Linear separators, iPython, maximal margin

W5 March 7-9: SVM, VC theorem and Bottou-Bousquet

W6 March 14-16: clustering (hierarchical, density based etc.), GMM, MRF, Apriori and association rules

W7 March 21-23: Recommender systems and generative models

W8 March 28-30: basics of neural networks, Sontag-Maas-Bartlett theorems, Bayes networks

W9 April 4-6:

W10 April 11-13: BN, CNN, MLP, Dropout, Batch normalization

W11 April 18-20: midterm, RNN

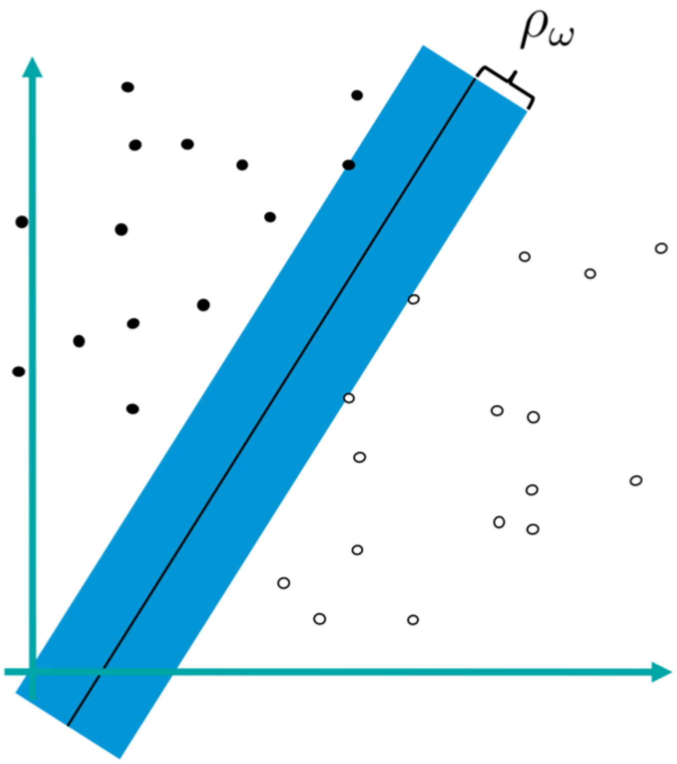
W12 April 25-27: LSTM, GRU, attention, Image caption, Turing Machine

W13 May 2-4: RBM, DBN, VAE, GAN

W14 May 9-11: Boosting, Time series

W14 May 16-18: TS, Projects on Friday

Maximal margin (recap)



Margin of an actual model:

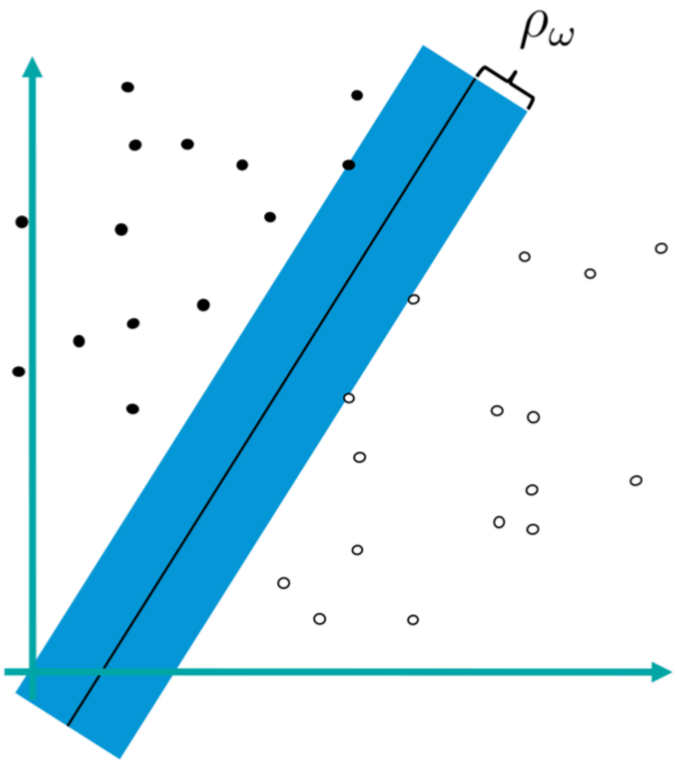
$$\rho_\omega(X) = \min_{x \in X} \frac{|x^T \omega|}{|| \omega ||}$$

The maximum margin problem is to maximize the margin while solving the original labeling problem:

$$\omega * = \arg\max_{\omega \in \Omega} \rho_\omega(X)$$

$$\text{subject to } y_t x_t^T \frac{\omega}{|| \omega ||} > 0, \forall t$$

Maximal margin (recap)



Margin of an actual model:

$$\rho_\omega(X) = \min_{x \in X} \frac{|x^T \omega|}{|| \ \omega \ ||}$$

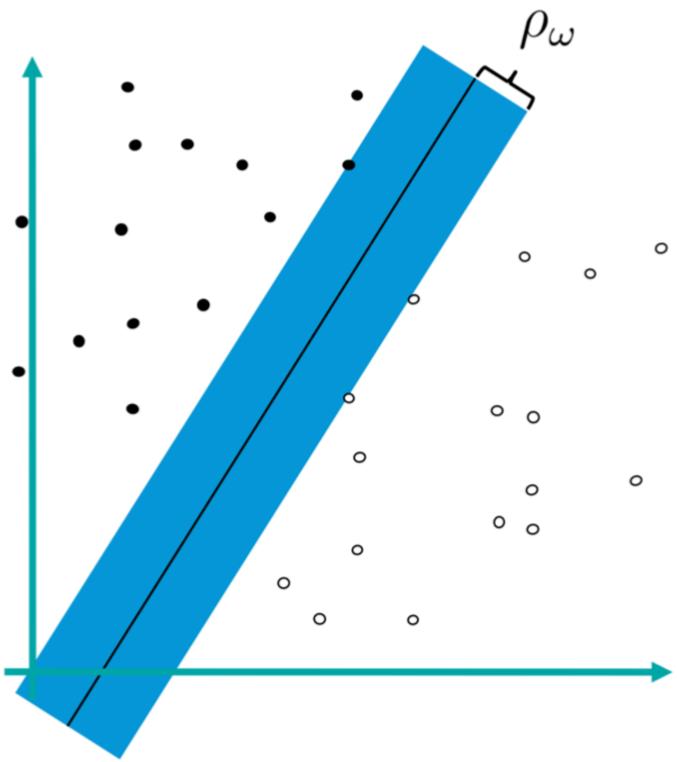
The maximum margin problem is to maximize the margin while solving the original labeling problem:

$$\omega* = \arg\max_{\omega \in \Omega} \rho_\omega(X)$$

$$\text{subject to } y_t x_t^T \frac{\omega}{|| \ \omega \ ||} > 0, \forall t$$

where the class labels are in {-1,+1}

Maximal margin



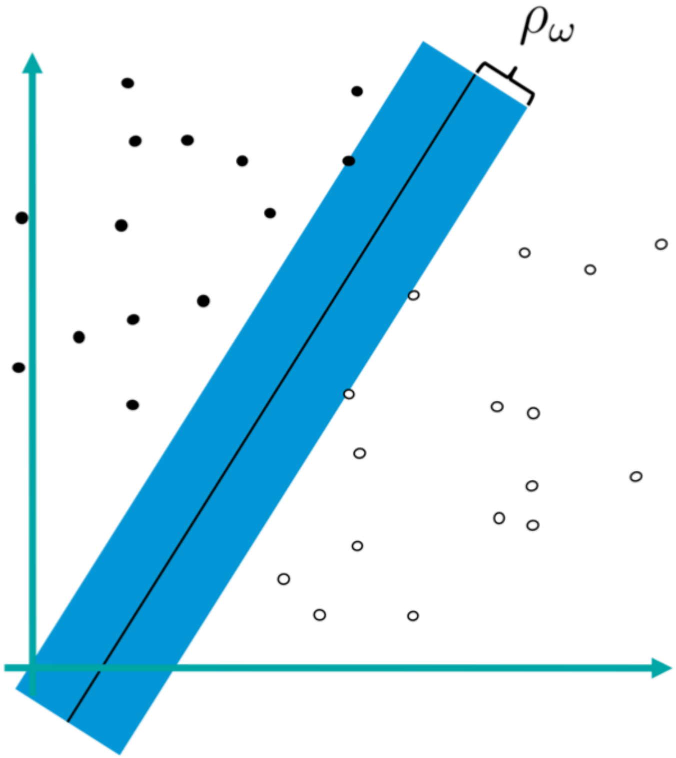
Because of the monotonicity of the sigmoid function, it can be expanded into a probabilistic sense:

$$\omega* = \arg\max_{\omega \in \Omega} \min_{x \in X} \mid p(x \mid \omega) - 0.5 \mid$$

$$\text{subject to } y_t x_t^T \frac{\omega}{\mid\mid \omega \mid\mid} > 0, \forall t$$

i.e. maximizing the minimum uncertainty (difference from the undecided probability).

Maximal margin



By definition (the minimum distance from the hyperplane (notice, we allow equality too))

$$y(x^T \frac{\omega}{||\omega||}) \geq \rho_\omega$$

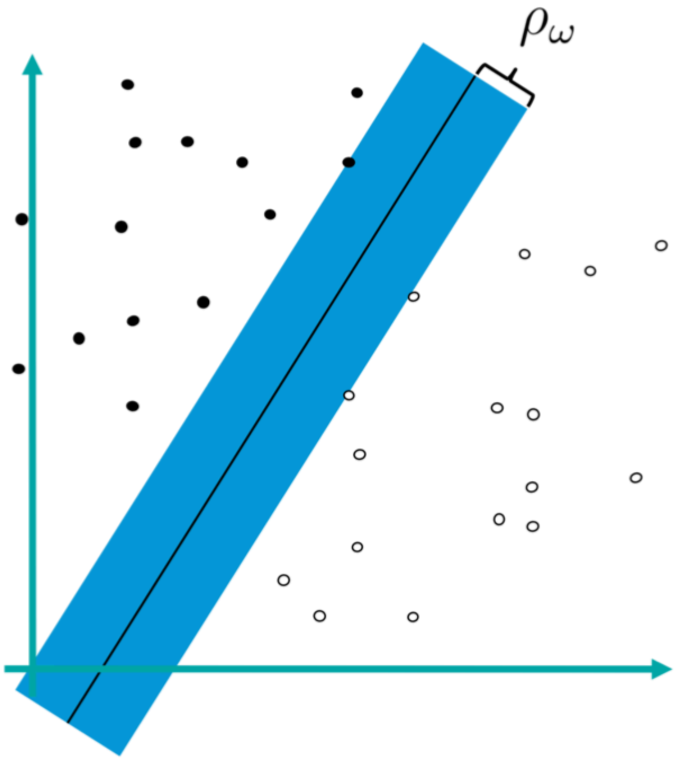
for all (x,y) pairs in the training set and so we can define a new hyperplane

$$\omega' = \frac{\omega}{||\omega||\rho_\omega}$$

for which the following holds for all (x,y):

$$y(x^T \omega') \geq 1$$

Maximal margin (still a bit recap)



The original maximization problem is equivalent to minimization of the norm of the new normal vector with a new constrain, formally

$$\text{minimize}_\omega \frac{1}{2} \mid\mid \omega \mid\mid^2$$

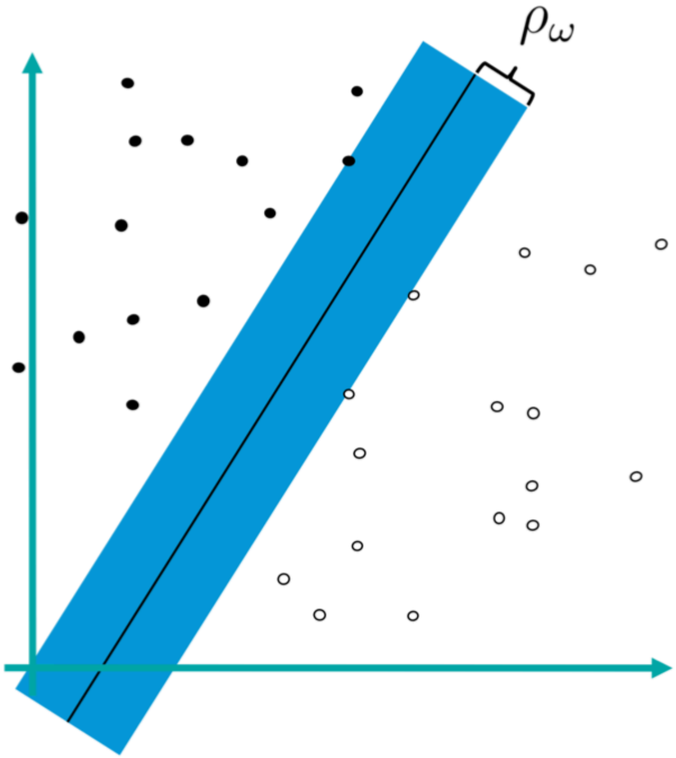
$$\text{subject to } y_t(x_t^T \omega) \geq 1, \forall t$$

Questions:

Why the ½ ?

Why the square?

Maximal margin (partially new stuff)



This convex, quadratic optimisation problem cannot be solved directly because of the constraints.

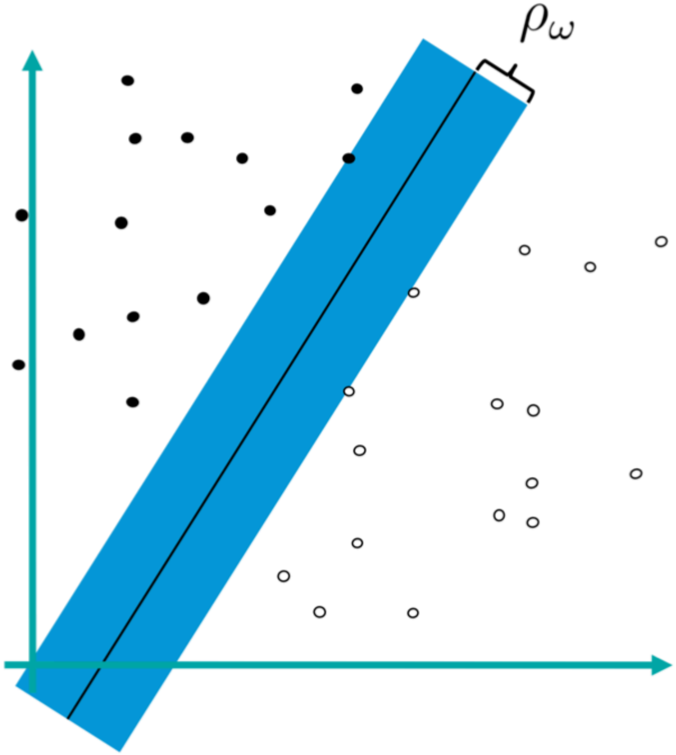
Both the constraint and value functions are cdf -> We can treat it is a Lagrangian problem

Formally, let be $\alpha_t \geq 0, \forall t$ the set of primal variables of the Lagrangian (multipliers).

Lagrangian function is

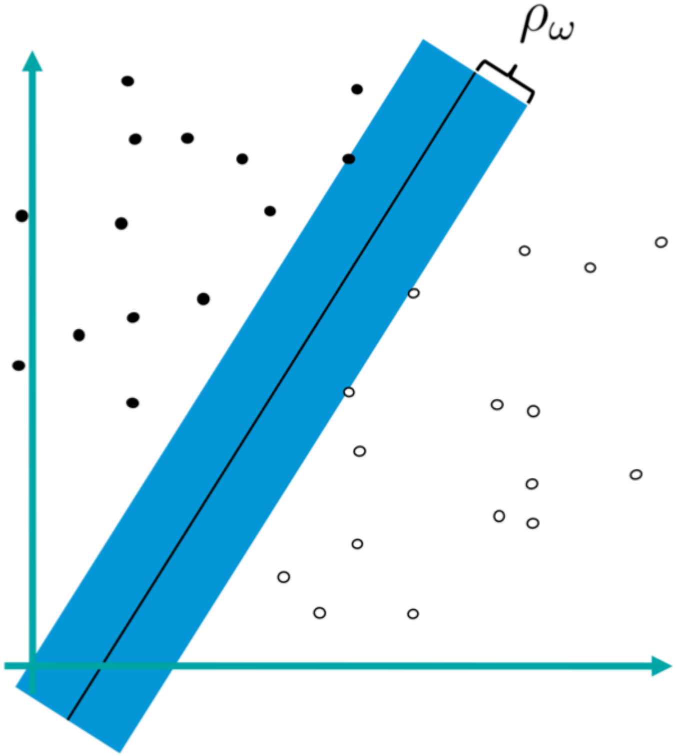
$$L(\omega, \alpha) = \frac{1}{2} \mid\mid \omega \mid\mid^2 - \sum_{t=1}^{T} \alpha_t (y_t (x_t^T \omega) - 1)$$

Maximal margin



Why Lagrange?

Maximal margin

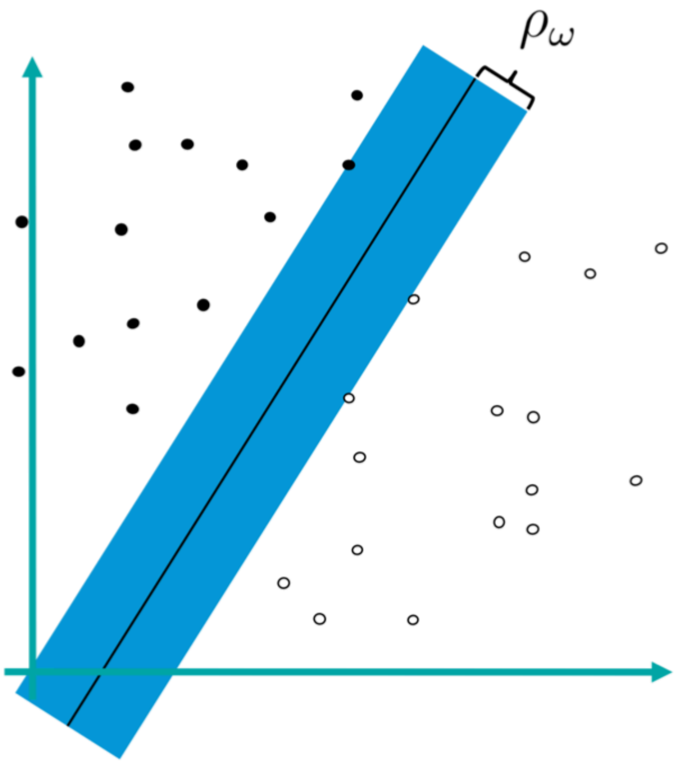


Why Lagrange?

The derivative respect to the normal vector will be zero at points where the original optimisation has usually an optimum (note, not all cases)

Let us derive the derivative!

Maximal margin



Why Lagrange?

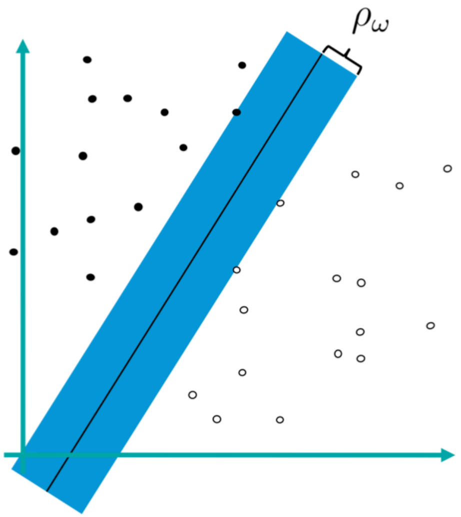
The derivative respect to the normal vector will be zero at points where the original optimisation has usually an optimum (note, not all cases)

Let us derive the derivative!

$$\frac{\partial L(\omega, \alpha)}{\partial \omega_i} = \omega_i - \sum_{t=1}^{T} y_t \alpha_t x_{ti} = 0$$

Stationary points: where the derivative is zero

Maximal margin



Recap:

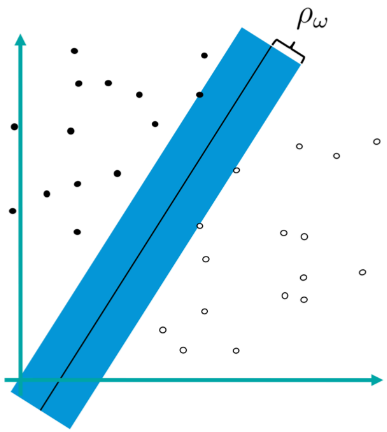
The normal vector is a linear combination of the training samples:

$$\omega = \sum_{t=1}^{T} \alpha_t y_t x_t$$

Let put everything into the kitchen sink:

$$L(\omega, \alpha) = L(\alpha) = \frac{1}{2}\sum_{i=1}^{T}\sum_{j=1}^{T} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{T} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{t=1}^{T} \alpha_t$$
$$= \sum_{t=1}^{T} \alpha_t - \frac{1}{2}\sum_{i=1}^{T}\sum_{j=1}^{T} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Maximal margin



So our final optimization (dual)

Notice, the second set of constraints come from the bias term.

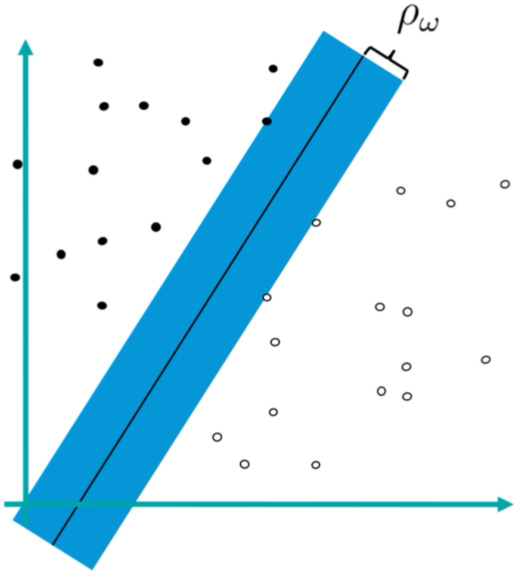
How can we solve such a problem? Seems complicated.

$$\text{maximize}_\alpha L(\alpha) = \sum_{t=1}^{T} \alpha_t - \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$subject\ to\ \alpha_i \geq 0, \forall i$$

$$\sum_{t=1}^{T} \alpha_t y_t = 0, \forall t$$

Maximal margin



Karush-Kuhn-Tucker conditions: actually two independent result:

Karush 1939 - master thesis
Kuhn-Tucker 1951 – independently

The optimal solution includes the positice Lagrangian multipliers and

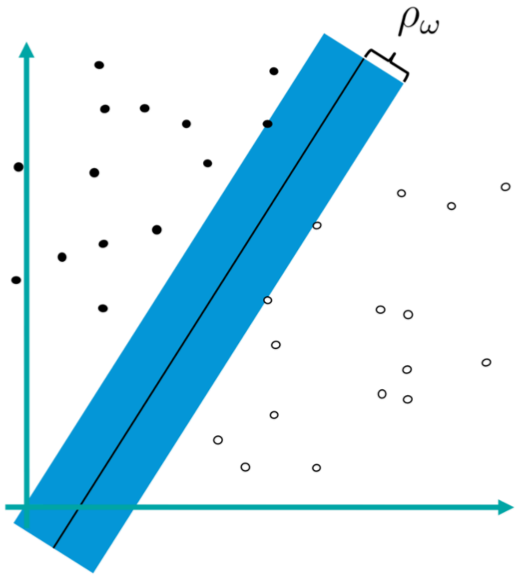
$$\alpha_i(y_i(x_i^T\omega)-1)=0,\forall i$$

What can be the interpretation?

Maximal margin

Karush-Kuhn-Tucker conditions: actually two independent result:

Karush 1939 - master thesis
Kuhn-Tucker 1951 – independently



The optimal solution includes the positice Lagrangian multipliers and

$$\alpha_i(y_i(x_i^T\omega)-1)=0, \forall i$$

What can be the interpretation?

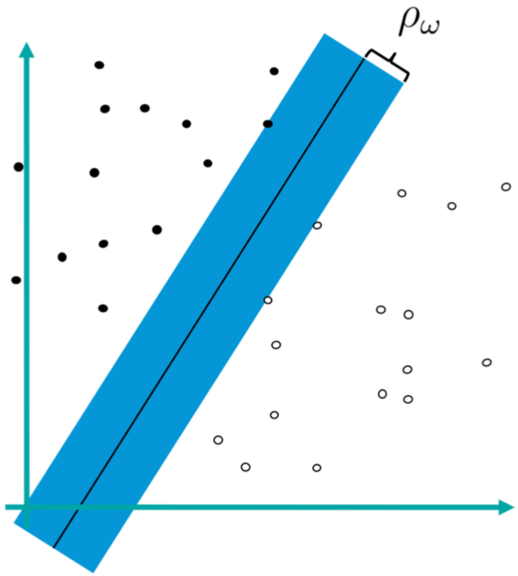
Cortes-Vapnik (1995) referred the training points with non-zero multipliers (aka positive according to KKT) as Support Vectors.

Maximal margin

The optimal solution includes the positice Lagrangian multipliers and

$$\alpha_i(y_i(x_i^T\omega) - 1) = 0, \forall i$$

What can be the interpretation?



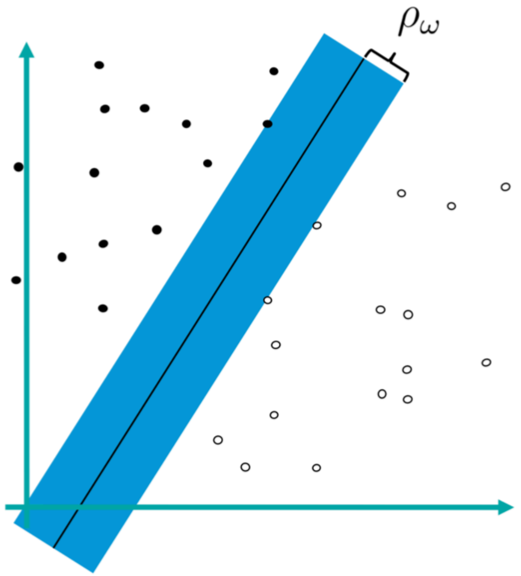
Cortes-Vapnik (1995) referred the training points with non-zero multipliers (aka positive according to KKT) as Support Vectors.

and

Unnecessary points?

$$\omega = \sum_{t=1}^{T} \alpha_t x_t = \sum_{x_i \in SV} \alpha_i x_i$$

Maximal margin (a bit of recap)



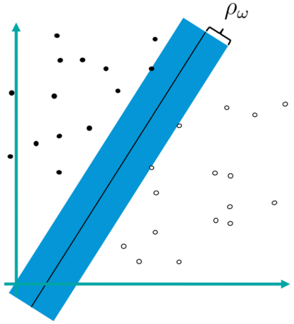
With a simple loss (penalty measure) the 1-Norm Soft Margin problem is

$$\text{minimize } \frac{1}{2} \| \omega \|^2 + C \sum_{t=1}^{T} \xi_i$$
$$\text{subject to } y_i(x_i^T \omega) \geq 1 - \xi_i, \forall i$$
$$\xi_i \geq 0, \forall i$$

where is the same as before (previously determined constant).

Let us check the Lagrangian!

Maximal margin (a bit of recap)



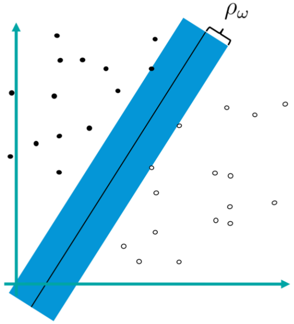
Lagrangian function:

$$L(\omega, \alpha, \beta) = \frac{1}{2} \mid\mid \omega \mid\mid^2 + C \sum_{t=1}^{T} \xi_i - \sum_{t=1}^{T} \alpha_t (y_t (x_t^T \omega) - 1 + \xi_t) - \sum_{t=1}^{T} \beta_t \xi_t$$

Beta is an additional set of Lagrange multipliers for the second constraint.

What is next?

Maximal margin (a bit of recap)



Lagrangian function:

$$L(\omega, \alpha, \beta) = \frac{1}{2} \mid\mid \omega \mid\mid^2 + C \sum_{t=1}^{T} \xi_i - \sum_{t=1}^{T} \alpha_t (y_t (x_t^T \omega) - 1 + \xi_t) - \sum_{t=1}^{T} \beta_t \xi_t$$

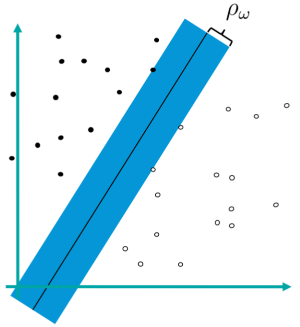
Beta is an additional set of Lagrange multipliers for the second constraint.

What is next? Derivative!

$$\frac{\partial L(\omega, \xi, \alpha, \beta)}{\partial \omega_i} = \omega_i - \sum_{t=1}^{T} \alpha_t y_t x_{ti} = 0$$

$$\frac{\partial L(\omega, \xi, \alpha, \beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0, \forall i$$

Maximal margin (a bit of recap)



Derivates:

$$\frac{\partial L(\omega,\xi,\alpha,\beta)}{\partial \omega_i} = \omega_i - \sum_{t=1}^{T} \alpha_t y_t x_{ti} = 0$$

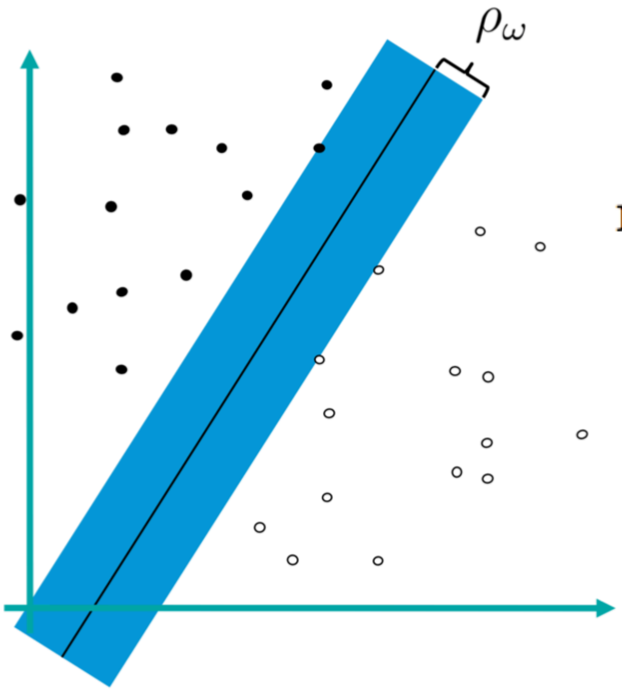
$$\frac{\partial L(\omega,\xi,\alpha,\beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0, \forall i$$

Notice the gradient respect to the normal vector does not include neither the loss or the second constraint -> identical to the case of non-soft margin!!! (Cortes-Vapnik 1995)

Since we know from KKT that both set of multipliers are positive:

$$\alpha_i(y_i x_i^T \omega - 1 + \xi_i) = 0, \forall i$$
$$\xi_i(C - \alpha_i) = 0, \forall i.$$

Maximal margin (a bit of recap)



The C upper is an upper bound and finally we arrived at:

$$\text{maximize } W(\alpha) = \sum_{t=1}^{T} \alpha_t - \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

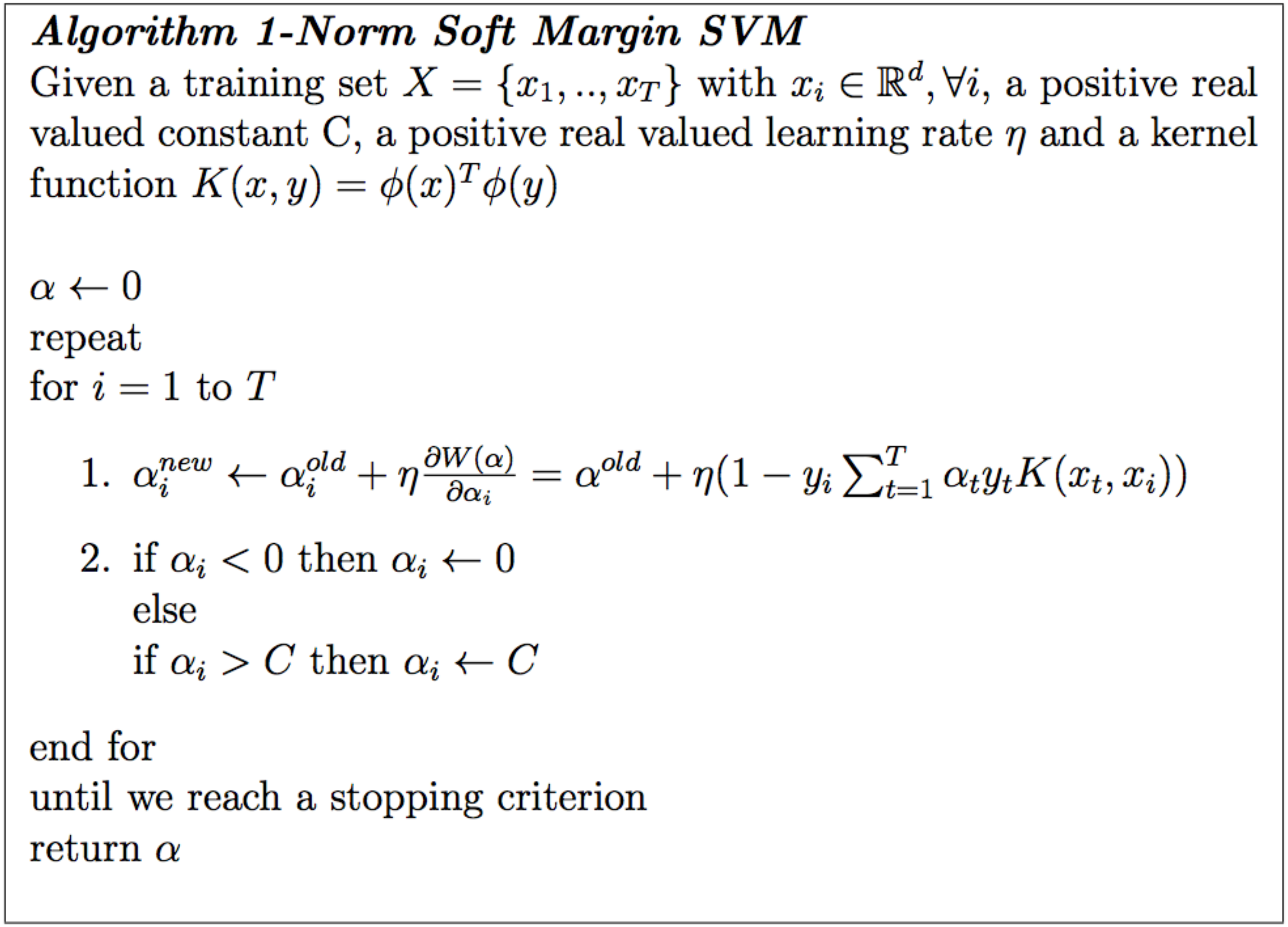
$$\text{subject to } 0 \leq \alpha_i \geq C, \forall i.$$

Where the derivatives are simple:

$$\frac{\partial W(\alpha)}{\partial \alpha_i} = 1 - y_i \sum_{t=1}^{T} \alpha_j y_j x_i^T x_j$$

Maximal margin (a bit of recap)

Replace the scalar product with a kernel and the final algorithm is:



Algorithm 1-Norm Soft Margin SVM

Given a training set $X = \{x_1, .., x_T\}$ with $x_i \in \mathbb{R}^d, \forall i$, a positive real valued constant C, a positive real valued learning rate η and a kernel function $K(x, y) = \phi(x)^T \phi(y)$

```
$\alpha \leftarrow 0$
repeat
for $i = 1$ to $T$

    1. $\alpha_i^{new} \leftarrow \alpha_i^{old} + \eta \frac{\partial W(\alpha)}{\partial \alpha_i} = \alpha^{old} + \eta(1 - y_i \sum_{t=1}^{T} \alpha_t y_t K(x_t, x_i))$

    2. if $\alpha_i < 0$ then $\alpha_i \leftarrow 0$
       else
       if $\alpha_i > C$ then $\alpha_i \leftarrow C$

end for
until we reach a stopping criterion
return $\alpha$
```

Brief intro to VC theorem

Lady (Dr. Muriel Bristol-Roach) :“by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup"

R. A. Fisher prepared 8 cups -> 4/4

He asked the Lady to choose 4 cups in which she thinks the milk was added first.

What is the probability of having 0,1,2,3,4 correct answers?

Brief intro to VC theorem

Lady (Dr. Muriel Bristol-Roach) :“by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup"

R. A. Fisher prepared 8 cups -> 4/4

He asked the Lady to choose 4 cups in which she thinks the milk was added first.

What is the probability of having 0,1,2,3,4 correct answers?

Overall: 70 cases

Out of them:

Having zero correct answer: 1
Having one correct answer: 16
Having two correct answers: 36
Having three correct answers: 16
Having four correct answers: 1

Vapnik-Chervonenkis theorem

The Vapnik-Chervonenkis theorem explains the connection between generalisation, training set selection and model selection.

Empirical risk:

$$R_{emp}(f) = \frac{1}{T}\sum_{t=1}^{T} l(f(x_i), y_i))$$

The theorem states that if we optimize for a binary loss function (0 if f(x_i) = y_i and 1 if not) over a set of independent samples from a fixed distribution D with known labels (the training set) than the true risk R_{true}(f) (the expected value of the loss function over D) is upper bounded by the empirical risk plus an additional value depending on the chosen function’s capabilities.

Vapnik-Chervonenkis theorem

The VC-theorem [Vapnik and Chervonenkis, 1971]:
the worst case scenario

For binary classification with a binary loss function and a chosen function class F the generalisation (the difference between the true and the empirical risk) is bounded as follows

$$P(\sup_{f \in \mathcal{F}} \mid R_{emp}(f) - R_{true}(f) \mid > \epsilon) \leq 8\mathcal{S}(\mathcal{F}, T)\mathrm{e}^{-\frac{T\epsilon^2}{32}}$$

and

$$\mathbf{E}[\sup_{f \in \mathcal{F}} \mid R_{emp}(f) - R_{true}(f) \mid|] \leq 2\sqrt{\frac{\log \mathcal{S}(\mathcal{F}, T) + \log 2}{T}}$$

Vapnik-Chervonenkis theorem

The theory shows that the bound is depending only on the size of the training set and the separating capability of the chosen function class measured by the shattering coefficient S(F,T), the maximum number of different labellings the function class F can realize over T samples.

Maximal number of labelings in case of binary classification?

S(F,T)=?

Vapnik-Chervonenkis theorem

The theory shows that the bound is depending only on the size of the training set and the separating capability of the chosen function class measured by the shattering coefficient S(F,T), the maximum number of different labellings the function class F can realize over T samples.

Maximal number of labelings in case of binary classification?

S(F,T)=?

For binary labels the maximum and the ideal would be $S(F,T) = 2^T$ but in practice usually it is not the case.

To capture this amount, they defined the so called Vapnik-Chervonenkis dimension (VC-dimension) that is independent from the size of the training set.

Vapnik-Chervonenkis theorem

The VC-dimension of a function class VC(F) is the cardinality of the largest set in the d-dimensional space which can be separated correctly (or shattered) with any label set.

According to Sauer's lemma [Sauer, 1972] the shattering coefficient is upper bounded

$$\mathcal{S}(\mathcal{F},T) \leq (1+T)^{VC(\mathcal{F})}$$

Linear separator?

Is it sharp bound?

Vapnik-Chervonenkis theorem

The VC-dimension of a function class VC(F) is the cardinality of the largest set in the d-dimensional space which can be separated correctly (or shattered) with any label set.

According to Sauer’s lemma [Sauer, 1972] the shattering coefficient is upper bounded

$$\mathcal{S}(\mathcal{F}, T) \leq (1 + T)^{VC(\mathcal{F})}$$

Linear separator?

Radon theorem (about convex sets): the VC-dimension of the linear separator (a hyperplane which separates the space into two half-spaces) is d + 1 in d-dimensional space

Is it sharp bound?

Vapnik-Chervonenkis theorem

The VC-dimension of a function class VC(F) is the cardinality of the largest set in the d-dimensional space which can be separated correctly (or shattered) with any label set.

According to Sauer’s lemma [Sauer, 1972] the shattering coefficient is upper bounded

$$\mathcal{S}(\mathcal{F},T) \leq (1+T)^{VC(\mathcal{F})}$$

Linear separator?

Radon theorem (about convex sets, we will discuss it later): the VC-dimension of the linear separator (a hyperplane which separates the space into two half-spaces) is d + 1 in d-dimensional space

Is it sharp bound?

No, imagine three points on a line in R^2

Vapnik-Chervonenkis theorem

E.g.

Let us consider a linear separator capable of separating with low empirical risk.

If the number of examples in the training set were high, the feature space may (!) had been high dimensional according to the theory.

This suggests a high shattering coefficient and high upper bound.

VC-dimension of the class of polynomial functions (kernel!) in R^d with degree D?

Vapnik-Chervonenkis theorem

E.g.

Let us consider a linear separator capable of separating with low empirical risk.

If the number of examples in the training set were high, the feature space may (!) had been high dimensional according to the theory.

This suggests a high shattering coefficient and high upper bound.

VC-dimension of the class of polynomial functions (kernel!) in R^d with degree D?

The embedded linear space has a dimension

$$d' = \sum_{k=1}^{D} \binom{d+k-1}{k} + 1$$

Vapnik-Chervonenkis theorem

Remember T is finite by definition, what are the consequences?

Vapnik-Chervonenkis theorem

Remember T is finite by definition, what are the consequences?

We can always find a polynomial function with a high enough degree where

$$d' > T+1$$

Empirical risk will be zero.

But what about the generalisation and the shattering coefficient?

Vapnik-Chervonenkis theorem

Remember T is finite by definition, what are the consequences?

We can always find a polynomial function with a high enough degree where

$$d' > T+1$$

Empirical risk will be zero.

But what about the generalisation and the shattering coefficient?

Shattering coefficient will be high if T is high -> the upper bound increase

☹️

Optimisation for low true risk is a balance between low empirical risk and low VC-dimension.

Vapnik Chervonenkis dimension in general (HK book)

Let U be a set of n points in the plane.

Let ε > 0 be a given error parameter.

Pick a random sample S of size s from U.

Let R be query rectangle, estimate |R ∩ U| by the quantity

$$\frac{n}{s}|R \cap S|$$

We wish to assert that the fractional error is at most ε for every rectangle R, i.e., that

$$\left||R \cap U| - \frac{n}{s}|R \cap S|\right| \leq \varepsilon n \ \text{ for every } \ R$$

Vapnik Chervonenkis dimension

There is a small probability that the sample is atypical: Example picking no points from a rectangle R which has a lot of points.

We can only assert the above with high probability or that its negation holds with very low probability:

$$\text{Prob}\left(\left||R\cap U| - \frac{n}{s}|R\cap S|\right| \; > \; \varepsilon n \;\; \text{for some } R\right) \leq \delta$$

where $\delta > 0$ is another error parameter.

Vapnik Chervonenkis dimension

Pick s samples uniformly at random from the n points in U. For one fixed R, the number of samples in R is a random variable which is the sum of s independent 0-1 random variables, each with probability $q = \frac{|R \cap U|}{n}$ of having value one. So, the distribution of |R∩S| is Binomial(s,q).

Using Chernoff bounds (chapter 2 in HK book!), for 0 ≤ ε ≤ 1:

$$\mathrm{Prob}\left(\left| |R \cap U| - \frac{n}{s} |R \cap S| \right| > \varepsilon n \right) \leq 2e^{-\varepsilon^2 s/(3q)} \leq 2e^{-\varepsilon^2 s/3}$$

Vapnik Chervonenkis dimension

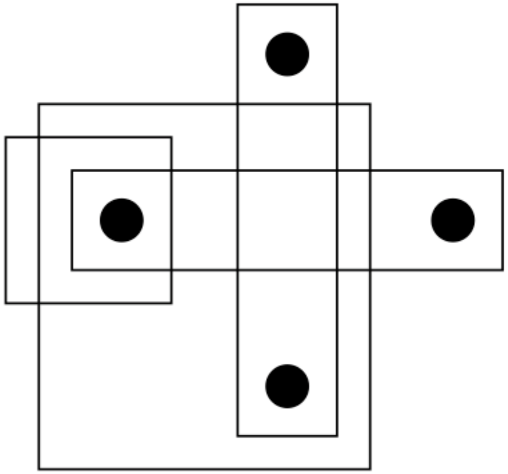
Set system (U,S) consists of a set U along with a collection S of subsets of U.

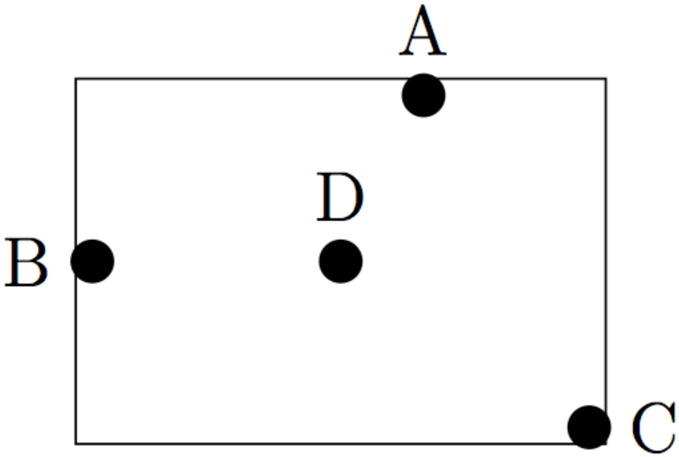
A subset A ⊆ U is shattered by S if each subset of A can be expressed as the intersection of an element of S with A.

The VC-dimension of the set system (U, S) is the maximum size of any subset of U shattered by S.

An example:
U = R^2 of points in the plane
S being the collection of all axis-parallel rectangles.

Vapnik Chervonenkis dimension





Vapnik Chervonenkis dimension

Theorem (Radon):

Any set $S \subseteq R^d$ with $|S| \geq d + 2$, can be partitioned into two disjoint subsets A and B such that convex(A) ∩ convex(B) = ∅

VC dimension of Half spaces in d-dimensions:

Define a half space to be the set of all points on one side of a hyper plane, i.e., a set of the form $\{x|a \cdot x \geq a_0\}$.

Radon's theorem implies that half-spaces in d-dimensions do not shatter any set of d+2 points.

Vapnik Chervonenkis dimension

Sketch of proof:

Divide the set of d+2 points into sets A and B. Suppose that some half space separates A from B. Then the half space contains A and the complement of the half space contains B. This implies that the half space contains the convex hull of A and the complement of the half space contains the convex hull of B. Thus, convex(A) ∩ convex(B) = ∅ a contradiction.

The VC-dimension of half spaces is **d + 1**.

Vapnik Chervonenkis dimension

Spheres in d-dimensions

A sphere in d-dimensions is a set of points of the form $\{x|\ |x - x0| \leq r\}$. The VC- dimension of spheres is $d + 1$.

Convex polygons

Consider the system of all convex polygons in the plane. For any positive integer n, place n points on the unit circle. Any subset of the points are the vertices of a convex polygon. Clearly that polygon will not contain any of the points not in the subset. This shows that convex polygons can shatter arbitrarily large sets, so the VC-dimension is **infinite**.

Vapnik-Chervonenkis theorem

Summary:

1) Low the empirical risk
2) Function class with low shattering coefficient (low complexity)
3) Let us take a disjoint test set, then according to the proof

$$P(\sup_{f\in\mathcal{F}} \mid R_{emp}(f) - R_{true}(f) \mid > \epsilon) \leq 2P(\sup_{f\in\mathcal{F}} \mid R_{emp}(f) - R'_{emp}(f) \mid > \frac{\epsilon}{2})/ \quad (4)$$

If we evalute on a separate test set we have an upper bound :)

Limitations?

Vapnik-Chervonenkis theorem

Summary:

1) Low the empirical risk
2) Function class with low shattering coefficient (low complexity)
3) Let us take a disjoint test set, then according to the proof

$$P(\sup_{f\in\mathcal{F}} \mid R_{emp}(f) - R_{true}(f) \mid > \epsilon) \leq 2P(\sup_{f\in\mathcal{F}} \mid R_{emp}(f) - R'_{emp}(f) \mid > \frac{\epsilon}{2}) / \quad (4)$$

If we evalute on a separate test set we have an upper bound :)

Limitations?

1) Fixed distribution … (e.g.?)
2) Really high for complex models, says nothing… ☹ or…?
3) Are we even close to the optimal during optimisation?

Bottou and Bousquet (2007)

The error has three parts and they are additive:

$$\mathcal{E} \;=\; \mathbb{E}\left[E(f^*_{\mathcal{F}}) - E(f^*)\right] \;+\; \mathbb{E}\left[E(f_n) - E(f^*_{\mathcal{F}})\right] \;+\; \mathbb{E}\left[E(\tilde{f}_n) - E(f_n)\right]$$

Approximation error   Estimation error   Optimisation error

Linear separator?

Limitations of the training set?

Gradient descent?

Bottou and Bousquet (2007)

Gradient based optimisations:

Gradient Descent (GD)

$$w(t+1) \;=\; w(t) - \eta \frac{\partial C}{\partial w}(w(t)) \;=\; w(t) - \eta \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \ell(f_{w(t)}(x_i), y_i)$$

Second order Gradient Descent (2GD):

$$w(t+1) \;=\; w(t) - H^{-1} \frac{\partial C}{\partial w}(w(t)) \;=\; w(t) - \frac{1}{n} H^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial w} \ell(f_{w(t)}(x_i), y_i)$$

where H is the Hessian

Bottou and Bousquet (2007)

Gradient based optimisations:

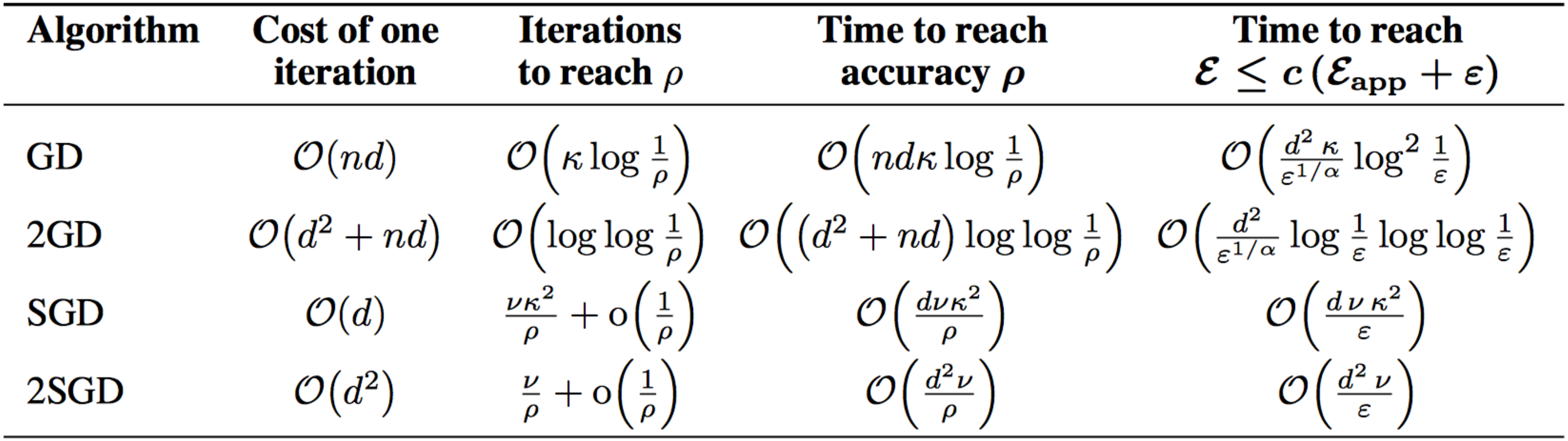
Stochastic Gradient Descent (GD)

$$w(t+1) \; = \; w(t) - \frac{\eta}{t} \, \frac{\partial}{\partial w} \ell\big(f_{w(t)}(x_t), y_t\big)$$

Second order Stochastic Gradient Descent (2GD):

$$w(t+1) \; = \; w(t) - \frac{1}{t} \, H^{-1} \, \frac{\partial}{\partial w} \ell\big(f_{w(t)}(x_t), y_t\big)$$

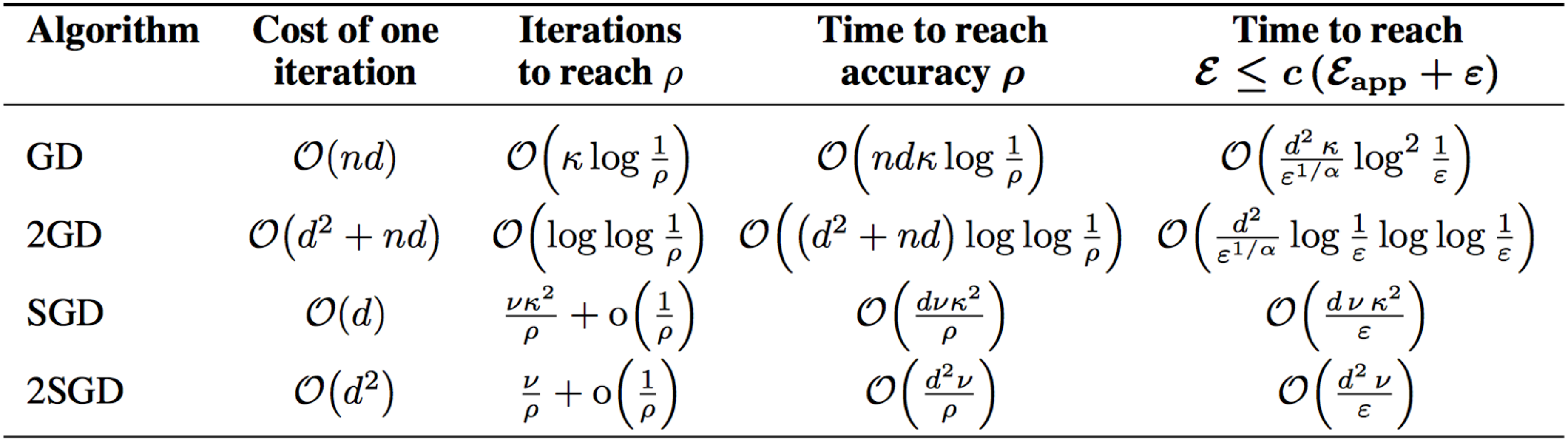
Bottou and Bousquet (2007)



Algorithm	Cost of one iteration	Iterations to reach ρ	Time to reach accuracy ρ	Time to reach $\mathcal{E} \leq c\,(\mathcal{E}_{\text{app}} + \varepsilon)$
GD	$\mathcal{O}(nd)$	$\mathcal{O}\left(\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(nd\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\,\kappa}{\varepsilon^{1/\alpha}} \log^2 \frac{1}{\varepsilon}\right)$
2GD	$\mathcal{O}\left(d^2 + nd\right)$	$\mathcal{O}\left(\log \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\left(d^2 + nd\right) \log \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2}{\varepsilon^{1/\alpha}} \log \frac{1}{\varepsilon} \log \log \frac{1}{\varepsilon}\right)$
SGD	$\mathcal{O}(d)$	$\frac{\nu\kappa^2}{\rho} + \mathrm{o}\left(\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d\nu\kappa^2}{\rho}\right)$	$\mathcal{O}\left(\frac{d\,\nu\,\kappa^2}{\varepsilon}\right)$
2SGD	$\mathcal{O}\left(d^2\right)$	$\frac{\nu}{\rho} + \mathrm{o}\left(\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\nu}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\,\nu}{\varepsilon}\right)$

Consequence?

Bottou and Bousquet (2007)



Algorithm	Cost of one iteration	Iterations to reach ρ	Time to reach accuracy ρ	Time to reach $\mathcal{E} \leq c\,(\mathcal{E}_{\mathrm{app}} + \varepsilon)$
GD	$\mathcal{O}(nd)$	$\mathcal{O}\left(\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(nd\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\,\kappa}{\varepsilon^{1/\alpha}} \log^2 \frac{1}{\varepsilon}\right)$
2GD	$\mathcal{O}\left(d^2 + nd\right)$	$\mathcal{O}\left(\log \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\left(d^2 + nd\right) \log \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2}{\varepsilon^{1/\alpha}} \log \frac{1}{\varepsilon} \log \log \frac{1}{\varepsilon}\right)$
SGD	$\mathcal{O}(d)$	$\frac{\nu\kappa^2}{\rho} + \mathrm{o}\left(\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d\nu\kappa^2}{\rho}\right)$	$\mathcal{O}\left(\frac{d\,\nu\,\kappa^2}{\varepsilon}\right)$
2SGD	$\mathcal{O}\left(d^2\right)$	$\frac{\nu}{\rho} + \mathrm{o}\left(\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\nu}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\,\nu}{\varepsilon}\right)$

Consequence?

Large-scale vs. small-scale learning act differently

And now for something completely different!

Project works