* A kernel will be valid if there exist a space such that

$$K(x_1, x_2) = \phi(x_1)^T \phi(x_2) = \sum_{n=1}^{N} \phi_n(x_1)\phi_n(x_2)$$

* consider a quadratic kernel with $D=2$, $x \cdot z$

$$K(x,z) = (x^T z)^2 = (x_1 z_1 + x_2 z_2)^2$$

$$= (x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2)$$

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

$x^T = (x_1 \ x_2)$

$x^T z = (x_1 \ x_2) \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

This can be expressed as an inner product of space where,

$$\phi(x) = x_1^2 + \sqrt{2} x_1 x_2 + x_2^2$$

$$\phi(z) = z_1^2 + \sqrt{2} z_1 z_2 + z_2^2$$

this gives, $\phi(x)\phi(z) = x_1^2 z_1^2 + 2 x_1 x_2 z_1 z_2 + x_2^2 z_2^2$

$$K(x,z) = \phi(x)^T \phi(z)$$

* A necessary and sufficient condition for a kernel function to be "valid" is that the gram matrix be positive and semidefinite for all choices of $\{\vec{x}_m\}$.

A Gram matrix of $\vec{x}$ is $x^T x$.

The linear vector $\vec{x}$ is projected into a quadratic surface.

If all the points in this surface are non-zero then our kernel is valid.

* Fisher's discriminant cost

$$J(w) = Tr\{(w S_W w^T)^{-1}\} \cdot (w S_B w^T)$$

— x — x — x — x $\boxed{\text{Tue Oct 31}}$

least square perceptron

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (h_n - t_n)^2$$ ← Does not work

min $E = 0$

max $E = \frac{N}{2}$

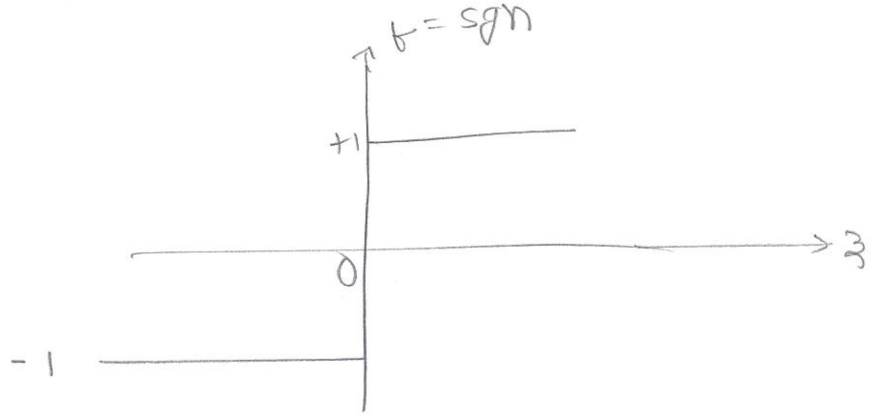
$0 \quad \frac{1}{2} \ldots \frac{N-2}{2} \quad \frac{N}{2}$ $\quad$ N+1 values

(N-2 mistake) (N mistakes)

we can not compute gradient, since function is discrete and not continuous.

cost = No. of misclassified patterns
= 0 for no mistake
$\frac{N}{2}$ for N mistakes

cost = $[0 \ \frac{1}{2} \ 2 \ldots \frac{N-2}{2} \ \frac{N}{2}]$ discrete set.



Distance metrics

① Euclidean $d(x,y) = \|x-y\|_2 = \sqrt{(x-y)^T (x-y)}$

② Hamming $d(x,y) =$ # of different values in fixed length strings

③ Mahalanobis distance $d(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}$

S is sample cov. matrix

$S = I \rightarrow$ Euclidean

$S = diag(\sigma_1^{-2}, \sigma_2^{-2}, \ldots) \rightarrow$ normalized Euclidean dist.

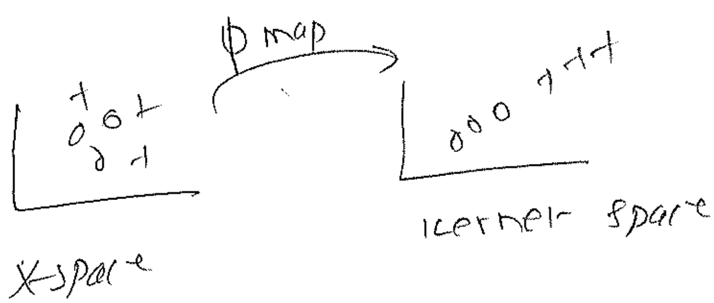
③ cosine similarity

$d(x,y) = 1 - \cos(x,y) = 1 - \frac{x^T y}{\|x\| \|y\|}$

④ Levenstein distance (edit distance)

min # of basis operations (del, insert, juxtapose) betn two strings

$x =$ 'athens' $y =$ 'hints' $d(x,y) = 4$



① SVM with slack (soft margin SVM)

$$\min J(w,b,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

with constraint

$$t_n(w^T\phi(x_n)+b) \geq 1-\xi_n$$

$$\xi_n \geq 0 \quad \text{and} \quad \sum_{i=1}^{N}\xi_i \leq Z \qquad \forall n \in \{1, \ldots$$

i.e. $1-\xi_n - t_n w^T\phi(x_n) - bt_n \leq 0$ —— ① (convex fn)

(convex fn constraint)

$-\xi_n \leq 0$ —— ②

$\forall n = 1,2,\ldots,N$

primal Lagrangian

$$L_p = L(w,b,\xi,\alpha,r) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\alpha_n(1-\xi_n-t_n w^T\phi_n - bt_n) - \sum_{n=1}^{N}r_n\xi_n$$

dual Lagrangian

$$L_D(\alpha,r) = \inf_{w,b,\xi} L_p(w,b,\xi,\alpha,r)$$

Now, $0 = \frac{\partial L_p(w,b,\xi,\alpha,r)}{\partial w} = w - \sum_n \alpha_n t_n \phi(x_n) \Rightarrow \boxed{w = \sum_n \alpha_n t_n \phi_n}$

$0 = \frac{\partial L_p}{\partial b} = -\sum_n \alpha_n t_n \Rightarrow \boxed{\sum_n \alpha_n t_n = 0}$

NO summation!

& $0 = \frac{\partial L_p}{\partial \xi_n} = C - \alpha_n - r_n \Rightarrow \boxed{C = \alpha_n + r_n} \quad \forall n \in \{1,2,\ldots,N\}$

Then,

$$L_D(\alpha, r) = \frac{1}{2}\|w\|^2 + C\sum_n \xi_n + \sum_n \alpha_n(1 - \xi_n - t_n w^T \phi_n - t_n b) - \sum_n r_n \xi_n$$

$$L_D = \sum_n \alpha_n - \frac{1}{2}\sum_{m,n=1}^{N} \alpha_m \alpha_n t_m t_n \, k(x_m, x_n)$$

Then the optimization problem in dual space is,

$$\underset{\alpha}{\text{maximize}} \quad L_D(\alpha) = \sum_n \alpha_n - \frac{1}{2}\sum_{m,n} \alpha_m \alpha_n t_m t_n \, k(x_m, x_n)$$

with constraints $\quad 0 \le \alpha_n \le C \quad \forall n = 1, 2, \dots, N$

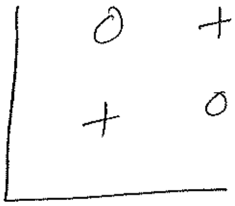
$$\sum_{n=1}^{N} \alpha_n t_n = 0$$

note: $\quad C\sum_n \xi_n - \sum_n \alpha_n \xi_n - \sum_n r_n \xi_n = 0$

since, $\quad C\sum_n \xi_n = \sum_n \alpha_n \xi_n + \sum_n r_n \xi_n$

$$= \sum_n (\alpha_n + r_n)\xi_n$$

(1)



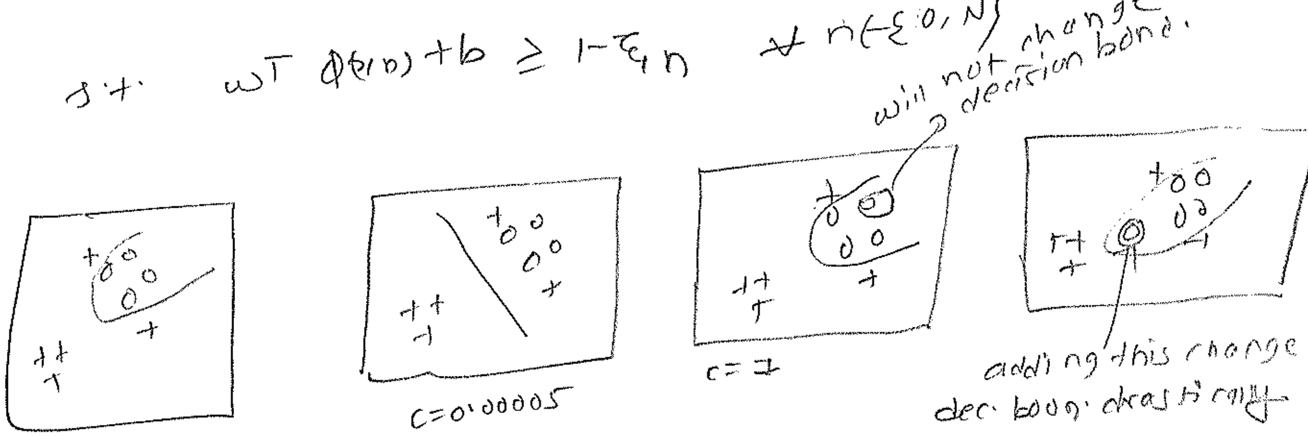
- SVM (quad kernel) can achieve zero training error
- Logis Regr, 3-NN cannot

(2) If examples are iid, increase training examples → may increase train error, ↳ but decrease test error

(3) SVM effect of 'C'

$$\min J(w,b) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad w^T\phi(x_n)+b \geq 1-\xi_n \quad \forall n \in \{0,N\}$$



C= 10,000 | C=0.00005 | C=1

between C>>1 and C≈0 choose C≈0 because it maximizes the margin between dominant cloud of points and we can not depend on any few data points which can be noise.

(4) Bias variance Tradeoff

	Bias	var
linear regr	high	low
d=2 poly	low	low
d=10 poly	low	high

python

① $X = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ append ones column to data X.

```
X = np.array([ [1,2,3], [4,5,6] ]) = np.arange(1,7).reshape(2,3)
ones = np.ones( X.shape[0] ).reshape(-1,1)
X1 = np.append( ones, X, axis=1 ).astype(np.int)
       concatenate
X1 = np.c_[ np.ones( X.shape[0] ).reshape(-1,1) , X ]
X1 = np.column_stack(( " ))
X1 = np.c_[ np.ones( X.shape[0] )[:, np.newaxis ] , X ]
                                  .reshape(-1,1)
   = np.c_[ np.atleast_2d( np.ones( X.shape[0] ) ).T , X ]
   = np.c_[ np.expand_dims( np.ones( X.shape[0] , axis=1 ) , X ]
   = np.c_[ np.
```

(5) Given $\phi(x) = [1, x_1, x_2, x_1 x_2]$

find the kernel $k(x, x')$.

$\Rightarrow$ $k(x, x') = 1 + x_1 x_1' + x_2 x_2' + x_1 x_2 x_1' x_2'$

(6) L1 vs L2 loss

(a) False: L2 is more robust to outlier than L1

L2 is more robust to outlier than L1 ... L2 loss can grow without bounds, but gradient of L2 loss can grow without bounds, but gradient of L1 loss is bounded, hence influence of outlier is limited.

(Note: L2 gives more values to misclassification than L1)

(b) L1 gives sparse solution & used in feature selection.

(c) Logistic loss is better than L2 loss in classification task.

(7) SVM small C,

For linearly separable data, small C can affect training accuracy.

A small C can allow large slacks, thus, the resulting classifier will have smaller w^2 and can have non-zero training error.

① pg ① solve the SVM problem without slack using Lagrange multiplier method

The optimization problem is

minimize $J(w,b) = \frac{1}{2}\|w\|^2$

s.t. $t_n(w^T\phi(x_n)+b) \geq 1 \quad \forall n \in \{1,\dots,N\}$

$1 \leq t_n(w^T\phi(x_n)+b)$

$1 \leq t_n w^T\phi(x_n) + t_n b$

$\underline{1 - t_n w^T\phi(x_n) - t_n b \leq 0}$ (convex fn constraint)

compare $f_i(x) \leq 0 \quad i = 1,\dots,m$

primal Lagrangian,

$L_p(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{n=1}^{N}\alpha_n(1 - t_n w^T\phi_n - t_n b)$

where $\alpha_n \geq 0$ are Lagrange multipliers.

Dual Lagrangian,

$L_D(\alpha) = \inf_{w,b} L_p(w,b,\alpha)$

first find the infimum of L_p w.r.t w,b:

$\frac{\partial}{\partial w}L_p = 0 = w + \sum_n \alpha_n(-)t_n\phi_n \Rightarrow \boxed{w = \sum_n \alpha_n t_n \phi_n}$ — ①

$\frac{\partial}{\partial b}L_p = 0 = \sum_n (-)\alpha_n t_n \Rightarrow \boxed{\sum_n \alpha_n t_n = 0}$ — ②

LOOK LHS

Then Dual Lagrangian is,

(2) $L_D(\alpha) = \frac{1}{2}\sum_{n,m} \alpha_m \alpha_n t_m t_n \phi_m \phi_n + \sum_n \alpha_n - \sum_n \alpha_n t_n \phi_n \cdot \sum_m \alpha_m t_m \phi_m$

$- \cancel{\sum_n \alpha_n t_n b}^{=0}$

$$\boxed{L_D(\alpha) = \sum_n \alpha_n - \frac{1}{2}\sum_{n,m} \alpha_m \alpha_n t_m t_n K(x_n, x_m)}$$

$K(x,y) = \vec{\phi}(x) \cdot \vec{\phi}(y) = \phi^T(x)\, \phi(y)$

Then the optimization problem in Dual space is,

maximize $L_D(\alpha) = \sum_n \alpha_n - \frac{1}{2}\sum_n \sum_m \alpha_m \alpha_n t_m t_n K(x_m, x_n)$

s.t. $\alpha_n \geq 0 \quad \forall n \in \{1 \cdots N\}$

$\sum_{n=1}^{N} \alpha_n t_n = 0$

Look here

Now, KKT conditions are,

note: we can write $y(x_n) = w^T\phi(x_n) + b$

(1) primal constraints $1 - t_n(w^T \phi(x_n) + b) \leq 0$ (convex constraint equation) $t_n(w^T\phi(x_n) + b) \geq 1$

(2) dual constraint $\alpha_n \geq 0 \quad \forall n \in \{1, \cdots N\}$

(3) commementary slackness $\alpha_n \{1 - t_n w^T \phi(x_n) - t_n b\} = 0$

for any data point, either $\alpha_n = 0$

$1 - t_n w^T \phi(x_n) - t_n b = 0$

these x_n are called support vectors

pg 3 here $n = 1, 2, \dots, N$

Suppose, out of N samples, there are m support vectors then $N-m$ samples will have Lagrange parameter $\alpha = 0$ and m examples will have non-zero Lagrange parameters.

$$1 - t_m w^T \phi(x_m) - t_m b = 0$$

or, $$1 - t_m \phi(x_m) \sum_n \alpha_n t_n \phi(x_n) - t_m b = 0$$

or, $$t_m b = 1 - t_m \phi(x_m) \sum_n \alpha_n t_n \phi(x_n)$$

$$= 1 - t_m \sum_n \alpha_n t_n \phi(x_n) \phi(x_m)$$

$$b t_m = 1 - t_m \sum_n \alpha_n t_n \phi(x_n) \phi(x_m)$$

$$b = \frac{1}{t_m} - \sum_n \alpha_n t_n \phi(x_n) \phi(x_m) = t_m - \sum_n \alpha_n t_n \phi(x_n) \cdot \phi(x_m)$$

$$\boxed{b = t_m - w \cdot \phi(x_m)}$$

$$\boxed{b = t_m - \sum_n \alpha_n t_n k(x_n, x_m)} \quad \because \frac{1}{t_m} = t_m$$

$$1 = \frac{1}{1} \text{ and } -1 = \frac{1}{-1}$$

this is true for all the m examples which have non zero Lagrange parameter α.

For numerical stability we choose value of b as the mean of all b-values, then,

$$\boxed{b = \frac{1}{|S|} \sum_{m \in S} \left[t_m - \sum_{n \in S} \alpha_n t_n k(x_n, x_m) \right]}$$

where S is the subset of all the examples where Lagrange parameter α is non zero.

$$S \subseteq D$$

$$S = \{ n \mid 1 - t_n w^T \phi(x_n) - t_n b = 0 \}$$

then, Linear discriminant function is,

$$\boxed{y(x) = w^T \phi(x) + b = \sum_{m \in S} \alpha_m t_m k(x, x_m) + \frac{1}{|S|} \sum_{m \in S} \left[t_m - \sum_{n \in S} \alpha_n t_n k(x_n, x_m) \right]}$$

KNN = memory-based (no model to fit)

① $\begin{cases} \text{find k nearest examples } x_1, x_2 \ldots x_k \text{ from test } x. \\ y(x) = \underset{t \in T}{\arg\max} \sum_{i=1}^{k} \delta_t(t_i) \end{cases}$

distance-weighted k-NN

w_i gives

$w_i = \frac{1}{\|x - x_i\|^2}$

KNN usage

① satellite image classification

② digit recognition

② mahalanobis dis

$d(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}$

sample covariance matrix

if $S = I$ = Euclidean distance

$S = diag(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_k^{-2})$

normalized Euclidean

③ cosine similarity

$d(x,y) = 1 - \frac{x^T y}{\|x\| \, \|y\|}$

④ kernel based distance weighted NN

→ binary classification $T \in \{+1, -1\}$

→ $y(x) = sign\left(\sum_{i=1}^{N} k(x, x_i) \cdot t_i\right)$

Wrapper method

① Greedy Forward selection

F = all features

S = subset of features

start S = { } empty

for each feature f in F − S

find best performing feature f^* and add to S.

Repeat until performance does not increase

or performance good enough

② Recursive Backward Elimination

F : {1, 2, ... K} is set of features

S = [ ] ranked set of features

Repeat until F − S is empty

train ω using linear SVM and F − S

find feature f with minimum $|\omega_f|$

append f to S

Return S

④ distance-weighted KNN for regression

1. find k nearest points $x_1, x_2 \dots x_k$

2. $y(x) = \dfrac{\sum_{i=1}^{k} w_i t_i}{\sum_i w_i}$ where $w_i = \dfrac{1}{\|x-x_i\|^2}$

$K=N \rightarrow$ Shepard's method

$y(x) = \sum_{i=1}^{N} K(x,x_i)\, t_i \,/\, \sum_{i=1}^{N} K(x,x_i)$ (kernel-based dist weighted)

⑤ Regression with KNN

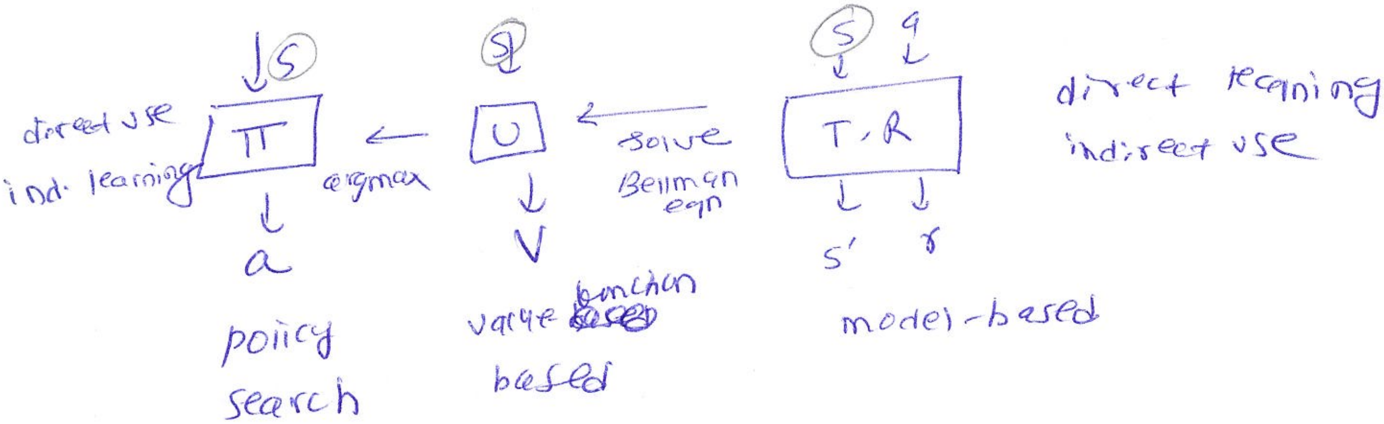
$y(x) = \frac{1}{K} \sum_{i=1}^{k} t_i$ (t_i are values, not classes)

⑥ distance weighted KNN (regression)

normal: $y(x) = \sum_{i=1}^{k} w_i t_i \,/\, \sum_{i=1}^{k} w_i$ $w_i = \dfrac{1}{\|x-x_i\|^2}$

kernel based: $y(x) = \sum_{i=1}^{N} K(x,x_i)\, t_i \,/\, \sum_{i=1}^{N} K(x,x_i)$ (t_i are values not classes)

3 Approaches to RL



Filter method of feature selection

① Mutual Information

$$MI(X,Y) = \sum_x \sum_y p(x,y) \ln \frac{p(x,y)}{p(x)\,p(y)}$$

=0 when x, y indep
max when x=y

Let there are K examples with l features.

②

$$\begin{array}{c|ccccc} & 1 & \cdots & j & \cdots & l \\ \hline 1 & & & & & \\ \vdots & & & O_{ij} & & \\ K & & & & & \end{array} \quad \} N_{x=i}$$

$\underbrace{\quad}_{N_{y=j}}$

O_{ij} = observed value for $x=i$, $y=j$

$$E_{ij} = \frac{N_{x=i}\, N_{y=j}}{N}$$

$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

classification Generative models

(3) probabilistic Generative

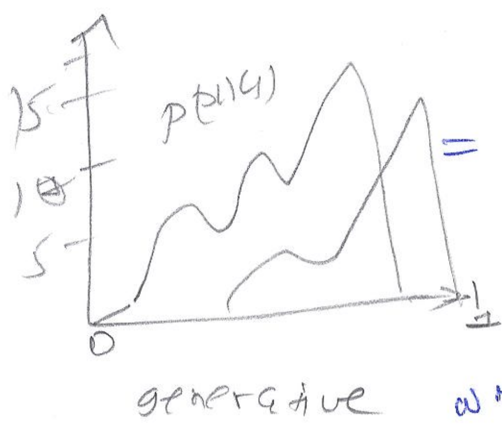
{ Naive Bias
Hidden Markov models

- inf. dec → separate
- can use $p(x)$ for outlier or novelty detection
- need to model dependencies between features

Naive Bias → resilient to noise

- text classification with NB → look HW
- multiple class
  posterior prob of class c_k given ~~resp~~ data x is,

$$p(c_k|x) = \frac{p(x|c_k) \cdot p(c_k)}{\sum_j p(x|c_j)\, p(c_j)}$$



$$= \frac{\exp(a_k(x))}{\sum_j \exp(a_j(x))}$$

normalized exponential
(softmax fn)

generative

where $a_k(x) = \ln p(x|c_k) \cdot p(c_k)$

① SVM for ranking

optimization problem

minimize $J(w, b) = \frac{1}{2} \|w\|^2 + C \sum \xi_{k,i,j}$

s.t. $w^T \phi(q_k, d_i) \geq w^T \phi(q_k, d_j) + 1 - \xi_{k,i,j}$

$\xi_{k,i,j} \geq 0$

(for a query q_k we want document d_i be ranked higher than document q_j)

1 Add ones column

```
= np.c_[np.ones(X.shape[0]).reshape(-1,1), X]
X = np.c_[np.ones(X.shape[0])[:,np.newaxis].T, X]
```

2
```
correct = np.sum(y_pred == y_test)
accuracy = correct / len(y_pred)
```

⑤ Gradient Descent (GD or BGD) GD with momentum

vanilla GD $v^{\tau+1} = \eta \nabla J(w^{\tau})$ $v^{\tau+1} = \gamma v^{\tau} + \eta \nabla J(w^{\tau})$

$w^{\tau+1} = w^{\tau} - v^{\tau+1}$ $w^{\tau+1} = w^{\tau} - v^{\tau+1}$

Batch Gradient Descent (Lect01, p.24)

$$J(w) = \frac{1}{2N} \sum_{n=1}^{N} (h_n - t_n)^2$$

$$w^{\tau+1} = w^{\tau} - \eta \nabla J(w^{\tau})$$

$$= w^{\tau} - \eta \sum_{n=1}^{N} (h_n - t_n)\, x_n$$

$= w^{\tau} -$ learninRate $(h-t)$ @ X1

confusion matrix

Actual class

		cat	not cat	
predicted class	cat	TP	FP	$\tilde{P}$
	not cat	FN	TN	$\tilde{N}$
		P	N	

true

00 = TN
01 = FP
10 = FN
11 = TP

	1	0
1	TP	FP
0	FN	TN
	P	N

actual

	0	1	
0	TN	FN	$\tilde{N}$
1	FP	TP	$\tilde{P}$
	N	P	

→ precision $= \frac{TP}{TP+FP}$

recall $= \frac{TP}{P} = \frac{TP}{TP+FN}$

(hit rate)

$\frac{\text{predicted true}}{\text{actual true}}$

F1- score is the Harmonic mean of precision and recall.

$$F1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

$$= \frac{2\,\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

* prove that the number of elements in
x and y is also a kernel

$\Rightarrow$ $k(x, y) = |x \cap y|$ is a kernel.

— x — x — x — x —

contd (Nov 11)

$$w^T \phi(x) = \sum_{n=1}^{N} \alpha_n t_n K(x_n, x)$$

since $S = n$; $\alpha_n > 0$ (strict)
$1 = 1, 2, \ldots N$
$n \notin S \Rightarrow \alpha_n = 0$

$$= \sum_{m \in S} \alpha_m t_m K(x_m, x)$$

$$+ \sum_{m \notin S} \alpha_m t_m K(x_m, x)$$

but if $x_m \notin S$, then $\alpha_m = 0$

* package libsvm or SVM lite,

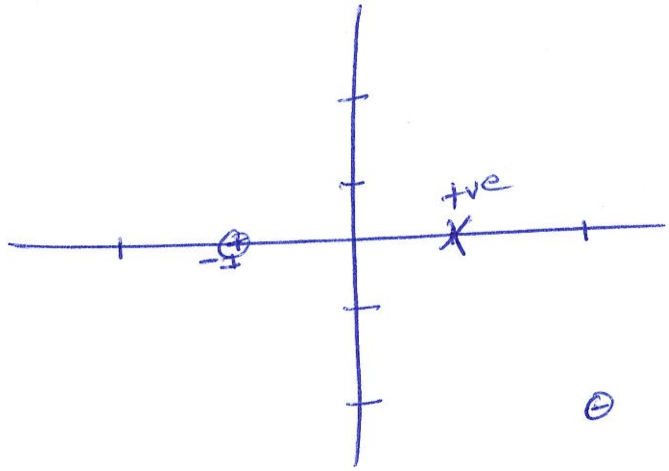
input $\rightarrow$ $\{(x_n, t_n)\}$ $1 \le n \le N$

output $\rightarrow$ $\{\alpha_m, t_m, x_m\}$ $m \in S$

model.txt $\rightarrow$ $\boxed{\begin{matrix} b \\ \alpha_m \; t_m \; x_m \end{matrix}}$

perceptron update

-1, 0 −ve
2, -2 −ve
1, 0 → +ve



final w = 0 1 2
let's say new point
$x = 2, -1.01$
$x_{aug} = 1\ 2\ -1.01$
$w x_{aug} = 0 + 2 - 2.02 < 0$
∴ −ve label

Augmented data

x_{aug}			t
1	-1	0	-1
1	2	-2	-1
1	1	0	1

initial wt $\vec{w} = (0\ 0\ 0)$

note we can also use $\eta\ \vec{w} \cdot \vec{x} + b$ original unaugmented x etc
$\eta = 0.9$
$b = 1$

	Test point	hypothesis correct classified?	Updated weights	correct example
eg1	−: (1 -1 0)	$wx = 0+0+0 < 0$ false	$w = w - x = (-1\ 1\ 0)$	
eg2	−: (1 2 -2)	$wx = -1+2+0 < 0$ false	$w = w - x = (-2\ -1\ 2)$	
eg3	+: (1 1 0)	$wx = -2-1+0 \geq 0$ false	$w = w + x = (-1\ 0\ 2)$	
eg1	−: (1 -1 0)	$wx = -1+0+0 < 0$ true	$w = w = (-1, 0\ 2)$	
eg2	−: (1 2 -2)	$wx = -1+0+(-4) < 0$ true	$w = w = (-1, 0, 2)$	
eg3	+: (1 1 0)	$wx = -1+0+0 \geq 0$ false	$w = w + x = (0\ 1\ 2)$	
eg1	−: (1 -1 0)	$wx = 0+(-1)+0 < 0$ true	$w = w = (0\ 1\ 2)$	
eg2	−: (1 2 -2)	$wx = 0+2+(-4) < 0$ true	$w = w = (0\ 1\ 2)$	
eg3	+: (1 1 0)	$wx = 0+1+0 \geq 0$ true	$w = w = 0, 1, 2$	

final weight

① constrained optimization
(Lagrange multipliers)

maximize $5-(x_1-2)^2-2(x_2-1)^2$

s.t. $x_1+4x_2=3$

soln: If we ignore constraint we get $x_1=2, x_2=1$
then $x_1+4x_2=2+4\cdot 1=6$ is too large
for the constraint.
consider
$L=L(x_1,x_2,\lambda) = 5-(x_1-2)^2-2(x_2-1)^2+\lambda(3-x_1-4x_2)$

look λ: $\lambda=0$ 2,1
$\lambda=1$ 3/2, 0
$\lambda=\frac{2}{3}\left(\frac{5}{3}, \frac{1}{3}\right)$ $\frac{5}{3}+4\cdot\frac{1}{3}=\frac{5}{3}+\frac{4}{3}=\frac{9}{3}=3$

formal soln

$\frac{\partial L}{\partial x_1} = -2(x_1-2)-\lambda=0$
$=-2x_1+4-\lambda=0 \Rightarrow 2x_1=4-\lambda$

$2x_1=4-\lambda$
$=4-\frac{2}{3}$
$=\frac{4\cdot 3-2}{3}=\frac{10}{3}$
$x_1=\frac{5}{3}$

$\frac{\partial L}{\partial x_2} = -4(x_2-1)-4\lambda=0$
$=-4x_2+4-4\lambda=0$
$\Rightarrow 4x_2=4-4\lambda$
$x_2=1-\lambda$

$\frac{\partial L}{\partial \lambda} = 3-x_1-4x_2=0$
$\Rightarrow 6-2x_1-8x_2=0$
$\Rightarrow 6-4+\lambda-8(1-\lambda)=0$
$\Rightarrow 6-4+\lambda-8+8\lambda=0$

$-6+9\lambda=0$
$9\lambda=6$
$\lambda=2/3$

op1

$$\boxed{\begin{aligned} &\min_{w,b} \ \tfrac{1}{2}\|w\|^2 \\ &\text{s.t.} \ t_n(w^T\phi(x_n)+b) \geq 1 \end{aligned}}$$

solution $= w_1^*, b_1^*$

op2

$$\boxed{\begin{aligned} &\min_{w,b} \ \tfrac{1}{2}\|w\|^2 \\ &\text{s.t.} \ t_n(w^T\phi(x_n)+b) \geq \gamma \end{aligned}}$$

$\Updownarrow$ op3

$$\boxed{\begin{aligned} &\min_{w,b} \ \tfrac{1}{2}\left\|\tfrac{w}{\gamma}\right\|^2 \\ &\text{s.t.} \ t_n\left(\left(\tfrac{w}{\gamma}\right)^T\phi(x_n) + \tfrac{b}{\gamma}\right) \geq 1 \end{aligned}}$$

op3 has sa(lu)tion

$w_2^* = w_1^*\gamma$

$b_2^* = b_1^*\gamma$

$\Updownarrow$ op3

rename

$w' = w/\gamma$

$b' = b/\gamma$

$$\boxed{\begin{aligned} &\min_{w,b} \ \tfrac{1}{2}\|w'\|^2 \\ &\text{s.t.} \ t_n(w'\phi(x_n)+b') \geq 1 \end{aligned}}$$

Now, decision hyperplanes are,

$$H1 = \{x \mid w_1^{*T}\phi(x) + b_1^* = 0\}$$

$$H2 = \{x \mid w_2^*\phi(x) + b_2^* = 0\}$$

$$= \{x \mid \gamma w_1^*\phi(x) + \gamma b_1^* = 0\}$$

$$= \{x \mid w_1^*\phi(x) + b_1^* = 0\}$$

$H2 = H1$ q.e.d

③ KMCP Kernel multi-class perceptron

1 define $f(x) = \sum_{i,j} \alpha_{ij} \left[\phi(x_i, t_i)^T \phi(x,t) - \phi(x_i, c_j)^T \phi(x,t) \right]$

$f(x) = w^T x = \sum_n \alpha_n t_n x_n^T x = \sum_n \alpha_n t_n K(x_n, x)$

2 initialize dual params $\alpha_{ij} = 0$

3 for $i = 1, \dots n$

4 $\quad c_j = \underset{t \in T}{\operatorname{argmax}} \; f(x_i, t)$ $(h_n = \operatorname{sign}(f(x)))$

5 $\quad$ if $c_j \neq t_i$ then $(h_n \neq t_n)$

6 $\quad \alpha_{ij} = \alpha_{ij} + 1$ $(\alpha_n = \alpha_n + 1)$

Repeat

Testing: $t^* = \underset{t \in T}{\operatorname{argmax}} \; f(x,t)$ $(h(x) = \operatorname{sign}(f(x)))$

② MCP (concept of kernel comes from here)

initialize parameters $w = 0$

for $i = 1 \cdot N$

$\quad c_j = \underset{t \in T}{\operatorname{argmax}} \; w^T \phi(x_i, t)$

$\quad$ if $c_j \neq t_i$ then

$\quad w = w + \phi(x_i, t_i) - \phi(x_i, c_j)$

w is loop invariant and is the weighted average

$$w = \sum_{i,j} \alpha_{ij} \left(\phi(x_i, t_i) - \phi(x_i, c_j) \right)$$

$$f(x) = w^T \phi(x,t) = \sum_{i,j} \alpha_{ij} \left(\phi(x_i, t_i)^T \phi(x,t) - \phi(x_i, c_j)^T \phi(x,t) \right)$$

① KP

define $f(x) = w^T x = \sum_n \alpha_n t_n K(x_n, x)$

initialize dual parameter $\alpha_n = 0$

for $i = 1 \dots N$

$\quad h_n = \operatorname{sign}(f(x))$

$\quad$ if $h_n \neq t_i$ then

$\quad \alpha_n = \alpha_n + 1$

Repeat

TEST: $y(x) = \operatorname{sign}(f(x))$

$$MI(X,Y) = \sum_x \sum_y p(X,Y) \cdot \ln \frac{p(X,Y)}{p(X) \cdot p(Y)} = \begin{cases} 0 & \text{when } X, Y \text{ indep} \\ \max & \text{when } X = Y \end{cases}$$

① mutual information
- → works with nominal
- → biased towards high parity feature
- → may choose redundant feature

② chi-square

$$E_{ij} = \frac{N_{x=i} \cdot N_{y=j}}{N}$$

l features →

$$\begin{array}{c|ccccc} & 1 & \cdots & j & \cdots & l \\ i & \cdots & \cdots & O_{ij} & \cdots & \} N_{x=i} \\ K & & & \vdots & & \\ & & & N_{y=j} & & \end{array}$$

K examples

$$X_0^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

③ pearson corr. coeff

$$\rho(X,Y) = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

④ SNR

$$\mu(X, \overset{\text{binary}}{Y}) = \frac{|\mu_+ - \mu_-|}{\sigma_x + \sigma_-}$$

⑤ T-Test

$$T(X,Y) = \frac{|\mu_+ - \mu_-|}{\sqrt{\frac{\sigma_+^2}{\mu_+} + \frac{\sigma_-^2}{\mu_-}}}$$

Filter vs wrapper

Filter	wrapper
1 much faster since no need to train the model	• computationally expen.
2 use statistical method of evaluation	• uses cross-validation
3 might fail to find best subset	• finds best subset
4 less prone to overfitting	• more prone to overfitting

① SVM for regression

$$\min J(w,b) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n)$$

s.t. $t_n \le t_n(w^T\phi(x_n)+b) + \varepsilon + \xi_n$

$t_n \ge t_n(w^T\phi(x_n)+b) - \varepsilon - \hat{\xi}_n$

$\xi_n, \hat{\xi}_n \ge 0 \quad \forall\ 1 \le n \le N$

② SVM for ranking

$$\min J(w,b) = \frac{1}{2}\|w\|^2 + C\sum \xi_{k,i,j}$$

s.t. $w^T\phi(q_k, d_i) \ge w^T(\phi(q_k, d_j)) + 1 - \xi_{k,i,j}$

$\xi_{k,i,j} \ge 0$

MDP markov Decision process

① policy
argmax

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi \right]$$

(policy expectation)

$$\pi^* = \underset{a}{\operatorname{argmax}} \sum_{s'} T(s,a,s') \, U(s')$$

(best action policy)

R = Reward

$\gamma^t R$ = discounted Reward

② utility
expectation & max only

$$U(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi \; , \; s_0 = s \right]$$

(expectation)

$$U(s) = R(s) + \gamma \max_{a} \sum_{s'} T(s,a,s') \, U(s')$$ ← Bellman eqn

* Naive Bias

3 Boolean input vectors $x_1 \, x_2 \, x_3$ and output y

- \# of parameters $= 2m+1 = 2 \times 3 + 1 = 7$

$p(y=1)$
$p(x_1=1 \mid y=0)$ $\quad p(x_1=1 \mid y=1)$
$p(x_2=1 \mid y=0)$ $\quad p(x_2=1 \mid y=1)$
$p(x_3=1 \mid y=0)$ $\quad p(x_3=1 \mid y=1)$

- \# of parameters if no conditional independence $= 1 + 2(2^m - 1)$
$= 1 + 2(2^3 - 1)$
$= 1 + 2(8-1)$
$= 1 + 14 = 15$

Example

category label

Train	Doc	words (w_i)	class	Total
3 documents for C1	1	Chinese Beiging Chinese	c1	$n_1 = 8$
	2	Chinese Chinese Shanghai	c1	
	3	Chinese Macao	c1	
1 document for C2	4	Tokyo Japan Chinese	c2	$n_2 = 3$
			?	11 words, 6 unique
Test	5			
Total	$\lvert D\rvert = \lvert D_1\rvert + \lvert D_2\rvert = 3+1 = 4$			

1 Chinese ||||
2 Beiging 1
3 shanghai 1
4 macao 1
5 Tokyo 1
6 Japan 1

} ignore

① vocabulary, V = { Chinese, Beiging, shanghai, macao, Tokyo, Japan }

$|V| = 6$

②

category C1 = c

- prior $P(c_1) = \frac{|D_1|}{|D|} = \frac{3}{4}$
- # of words in class C1, $n_1 = 8$

conditional probabilities

$n_{ki} = n_{11} = 5$ → Laplace smoothing

$P(w_1|c_1) = P(\text{Chinese}|c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$ (n_k, $|V|$)

$P(w_2|c) = P(\text{Beiging}|c) = \frac{1+1}{14} = \frac{2}{7}$

$P(w_3|c_1) = P(\text{shanghai}|c) = 2/7 = P$

$P(w_2|c_1) = P(\text{Tokyo}|c) = \frac{0+1}{8+6} = \frac{1}{14}$

$P(w_3|c_1) = P(\text{Japan}|c) = \frac{0+1}{8+6} = \frac{1}{14}$

category C2 = j (japan)

- prior $P(c_2) = \frac{|D_2|}{|D|} = \frac{1}{4}$
- # of words in class C2, $n_2 = 3$

$P(w_1|c_2) = P(\text{Chinese}|j) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(w_2|c_2) = P(\text{Tokyo}|j) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(w_3|c_2) = P(\text{Japan}|j) = \frac{1+1}{3+6} = \frac{2}{9}$

③ choosing a class

prior, Chinese, Japan, Tokyo

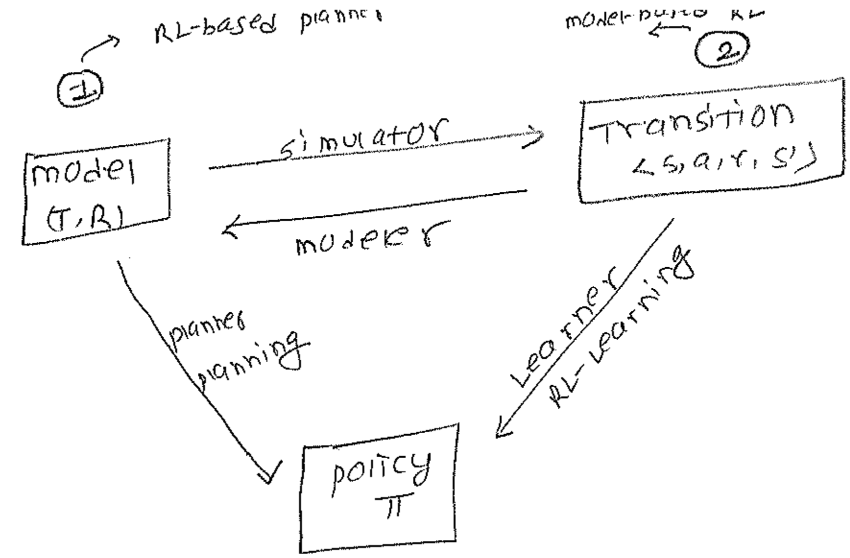
$$P(c_1|x) \propto P(c_1) \prod_i P(w_i|c_1) = \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14} \cong 0.0003$$

$$P(c_2|x) \propto P(c_2) \prod_i P(w_i|c_2) = \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} \approx 0.0001$$

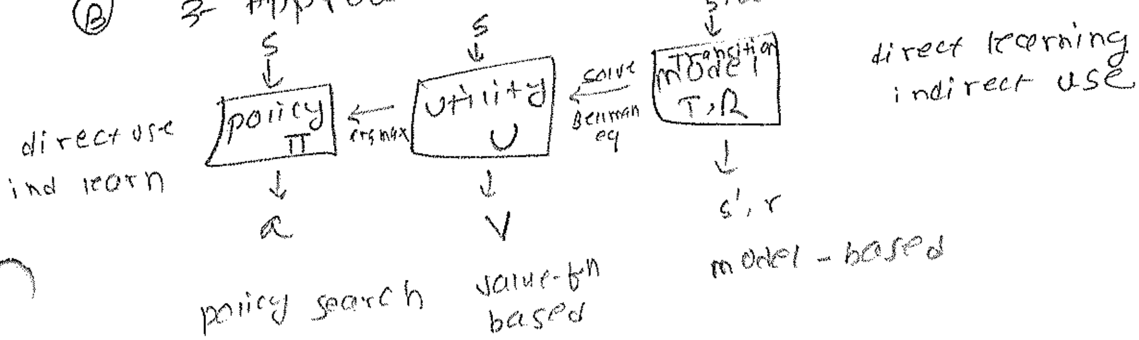
$$c^* = \underset{c_k}{\operatorname{argmax}}\ P(c_k) \prod_{j=1}^{n} P(w_j|c_k) = \underset{c_k}{\operatorname{argmax}} \{0.0003, 0.0001\}$$

= largest 0.0003 = C1

Ⓐ



Ⓑ 3 Approaches to RL



Ⓒ Q function

$$U(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s')\, U(s')$$ (Bellman eqn) utility is a scalar

change u to Q

$$\pi(s) = \arg\max_a \sum_{s'} T(s,a,s')\, U(s')$$ (policy gives an action)

$$Q(s,a) = R(s) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q(s',a')$$

quiz $\Rightarrow$ $U(s) = \max_a Q(s,a)$

$\Rightarrow$ $\pi(s) = \arg\max_a Q(s,a)$

HNB

① mutual information

$$MI(X,Y) = \sum_{x}\sum_{y} p(x,y) \cdot \ln\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \quad = 0 \text{ when } x,y \text{ indp}$$

$$= KL\left[p(x,y) \,\|\, p(x)\,p(y)\right] \quad = \max \text{ when } x = y$$

bad → biased towards high arity features

bad → may choose redundant feature

bad → works only with nominal features + labels

* 3 parametric Approaches to

① Discriminant function

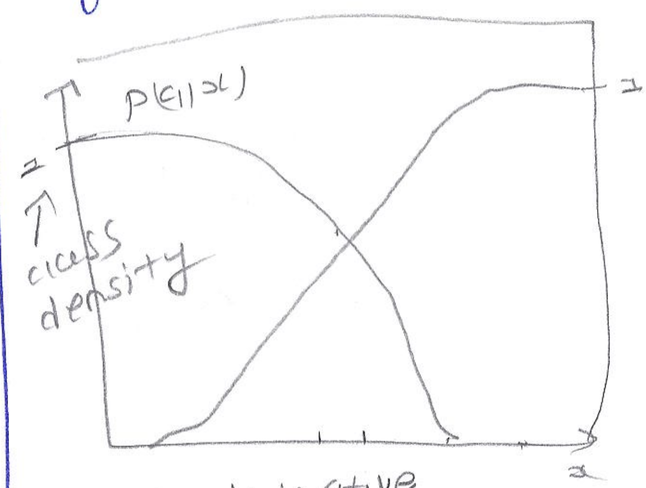
- Fisher's linear disc
- perceptron
- SVM

inference and decision are combined as single learning problem

② probabilistic discriminative models

- logistic regression
- conditional random fields

- inf, dec → separate
- less data need to compute $p(C_k|x)$ than $p(x|C_k)$
- can accommodate many overlapping features



discriminative

③ pearson corr coeff

$$\rho(X,Y) = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$ (population)

(sample)

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$-1 \le \rho(X,Y) \le 1$$

④ SNR

binary classes = $\{+,-\}$
mean = $\{\mu_+, \mu_-\}$

examples are binary
$y \in \{+1,-1\}$

$$\mu(X,Y) = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}$$

μ_+, σ_+ = mean & std for +ve class samples

③ T-test

$$T(X,Y) = \frac{|\mu_+ - \mu_-|}{\sqrt{\frac{\sigma_+^2}{\mu_+} + \frac{\sigma_-^2}{\mu_-}}}$$

CS 4900/5900: Machine Learning
Fall 2017

Class Meetings: Tue, Thu 10:30–11:50am, ARC 212
Instructor: Razvan Bunescu
Office: Stocker 341
Office Hours: Tue, Thu 12:00–12:30pm, or by email appointment
Email: bunescu @ ohio edu
Class Website: `http://ace.cs.ohio.edu/~razvan/courses/ml4900`

Prerequisites:

The students are expected to be comfortable with programming and familiar with basic concepts in linear algebra and statistics.

Textbook:

There is no textbook for this class. Slides and supplementary materials will be made available on the course website.

Supplementary Texts:

Machine Learning: The Art and Science of Algorithms that Make Sense of Data
by Peter Flach, Cambridge University Press, 2012
A Course in Machine Learning [free online]
by Hal Daume III
Machine Learning
by Tom Mitchell. McGraw Hill, 1997
Pattern Recognition and Machine Learning
by Christopher Bishop. Springer, 2007
Pattern Classification
by Richard O. Duda, Peter E. Hart, & David G. Stork. Wiley-IS, 2001
The Elements of Statistical Learning: Data Mining, Inference, and Prediction
by T. Hastie, R. Tibshirani, & J. H. Friedman. Springer Verlag, 2009

Course Description:

This course will give an overview of the main concepts, techniques, and algorithms underlying the theory and practice of machine learning. The course will cover the fundamental topics of classification, regression and clustering, and a number of corresponding learning models such as perceptrons, logistic regression, linear regression, Naive Bayes, nearest neighbors, and Support Vector Machines. The description of the formal properties of the algorithms will be supplemented with motivating applications in a wide range of areas including natural language processing, computer vision, bioinformatics, and music analysis. The topics covered in this course will also prepare students for taking more advanced courses in data mining and deep learning.

Grading:

50%: Homework Assignments
20%: Midterm Exam (Oct 12, in class) 1 hr 20 min
30%: Final Exam (Dec 12, 10:10am – 12:10pm)

Grading Scale:

A (> 92%) A–(> 90%) B+(> 87%) B(> 83%) B–(> 80%)
C+(> 77%) C(> 73%) C–(> 70%) D+(> 67%) D(> 63%) D–(> 60%)

Important Dates:

Friday, Sep 1: Last day to add class.
Tuesday, Oct 10: Reading Day, no class.
Friday, Nov 3: Last day to drop class.
Thursday, Nov 23: Thanksgiving break, no class.
Thursday, Dec 7: Last day of this class.

Course and Attendance policies:

Assignments: `All homework assignments are due before the class.` No late submissions will be accepted without prior approval.

Attendance: `It is in your best interest to attend the lectures.` Some of the material will not be found in the supplementary text or on the slides. Extra credit will be awarded for class activity. Also, be sure to check your OU email for important announcements on a regular basis.

Academic Dishonesty Policy:

`All work must be the student's own.` All external references used in reports must be properly cited. No credit will be given for duplicate or plagiarized work. Additional measures may be imposed by the University Judiciaries, when conditions warrant. Students may appeal academic sanctions through the grade appeal process. The OU Student Code of Conduct Policy is available online at:

`http://www.ohio.edu/communitystandards/academic/students.cfm`

Disability-based Accommodation:

Any student who suspects s/he may need an accommodation based on the impact of a disability should contact the class instructor privately to discuss the students specific needs and provide written documentation from the Office of Student Accessibility Services. If the student is not yet registered as a student with a disability, s/he should contact the Office of Student Accessibility Services.

Other Policies:

Be sure to notify the professor of any exam conflicts or other extenuating circumstances well in advance. `No missed exams will be made up without prior approval.` Medical excuse forms need to explicitly mention that the student could not have attended the exam at the specified time due to health concerns.