(P) Què significan els asteriscos generats amb l'executable en alguns valors de les yi?

(R) Mireu la transparència 75 del laboratori i trobareu la resposta a la pregunta.

(P) Per fer la validació hem d'utilitzar un nou conjunt de dades diferent al conjunt de trainning? I si és així ha de ser de la mateixa dimensió?

(R) Això feu-ho com vulgueu, podeu usar un conjunt diferent de la mateixa o diferent dimensió. Teniu llibertat per triar el que vulgueu. Jo estic més interessat en els aspectes propis d'optimització que en els d'usar la SVM.

(P) En el punt tres de l'esdeveniment d'atenea, diu que hem d'utilitzar un altre data set, et refereixes al mateix dataset de mides diferents o a un extret d'internet?

(R) A un altre diferent, buscat per internet.

(P) Quan utilitzem la funció: sklearn.datasets.make_swiss_roll(100) no entenem el segon array que ens retorna ni tampoc com s'extrauria la target.

(R) No esteu obligats a usar datasets generats per sklearn.datasets.make_swiss_roll(), podeu usar qualsevol dataset que doni dades linealment no-separables. Jo us vaig comentar la funció sklearn.datasets.make_swiss_roll() perquè genera dades d'aquest tipus, però no és obligatori usar-la. De totes formes si la voleu usar, el segon array (li dic t) que retorna, tal i com diu el help de la funció, és "The univariate position of the sample according to the main dimension of the points in the manifold". No sé exactament com es calcula aquest array t, però la idea es que, si desenrotllessis les dades (si les pinteu semblen un pastíl tipus "braç de gitano" enrotllat), la posició que tindrien a l'eix on s'ha desenrotllat es la que et dona t. Per tant podeu usar aquest valor per generar la target. Per exemple, valors per sota de la mitjana o mediana de t són d'una classe i per sobre són de l'altra
classe. Això ho vaig explicar a la classe de dijous 14 de maig.

(P) Problemes llegint les dades amb asteriscs en AMPL. S'ha de fer tot el projecte en AMPL o pot usar-se un altre llenguatge per tractar el fitxer de dades?(R) Les dades que genera el generador no són directament llegibles per AMPL, les heu

d'adaptar. En particular, heu d'eliminar els *, que només són informatius per a que sapigueu quins punts a priori no respectaran la regla de classificació. La transformació i/o adaptació de les dades la podeu fer com vulgueu (manualment perquè es senzill, o si voleu en python). AMPL només cal usar-lo per a l'optimització; la manipulació de les dades i anàlisi de resultats feu-ho com us vagi millor.

(P) Les dades del vector y amb un * les hem de canviar pel seu oposat? És a dir, hem de canviar els -1* per 1 i els 1* per -1?

(R) No, heu de mantenir el valor que tenen, així tindreu dades linealment no-separable. El * l'indico només per a que sabeu on hi haurà les dades mal classificades.

(P) Pel que fa a la funció kernel, com es calcula la variància?

(R) Podeu provar amb el valors que vulgueu. Si mireu https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html veureu que python usa gamma=1/n (n= n. features) (on gamma= 1/(2*sigma^2) en el nostre cas). Podeu usar aquest valor.

(P) Quan treballem amb les dades no separables, com calculem la w? No sabem com fer-ho, donat que no coneixem la funció fi(x).

(R) No es pot calcular, si no sabeu la funció de transformació phi(x), com passa amb el kernel RBF. En aquest cas no cal calcular (w,gamma), només us cal poder calcular w'*phi(x)+gamma per veure a quina classe pertany un punt x, i això ho podeu fer sense saber phi(x). Ho vaig explicar al darrer video del Tema 3 de dualitat.

(P) No sabem com llegir les dades. És a dir, al tenir totes les x i les y en una mateixa matriu no sabem com seleccionar les variables per separat.

(R) Si la variable y és a la 5a columna de la matriu, en AMPL accediu fent A[i,5] (per a l'element i). I així per a totes les dades.

(P) No sabem si hem de treure els * del data set o s'han de mantenir.

(R) Sí, s'han de treure per llegir les dades, només són indicatius (mireu les respostes a preguntes similars).

(P) Com sabem quin valor li hem de donar als paràmetres sigma (del kernel) i a la nu?

(R) Per sigma ho he explicat abans en una altra pregunta. Per nu, heu de provar, no hi ha un nombre màgic. Podeu mirar el manual de SVC de python i usar el mateix valor (allà es diu C, i acostuma a ser un nombre entre 1 i 10; però millor que proveu diferents valors i així entendreu com afecta a la
solució de la SVM).