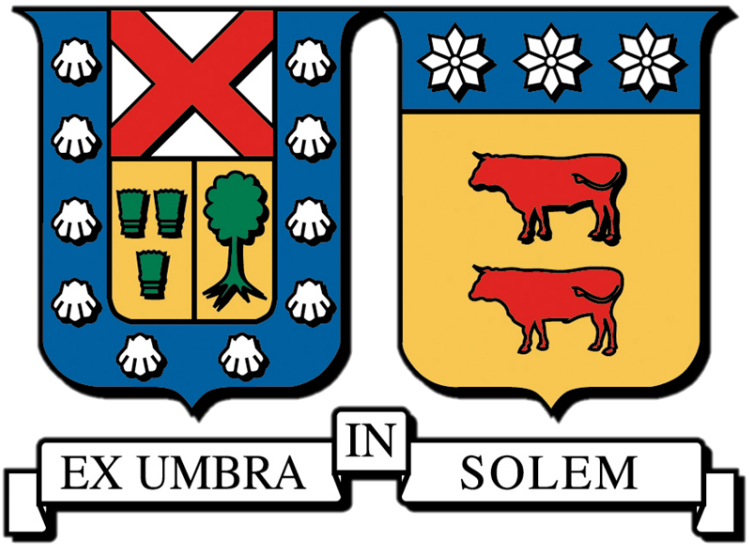

Universidad Técnica Federico Santa María
Proyecto Optimizacion No Lineal MAT-279

Técnicas de Optimización No Lineal Aplicadas al Diagnóstico de Cáncer Mamario

Hugo Andrés Rocha Alfaro	Rol:201610531-K
Mario Carlos Mallea Ruz	Rol:201710515-1
Maximiliano Ramírez Núñez	Rol:201710507-0

Segundo Semestre 2020

Índice general

0.1. Introducción

En el presente informe se estudiará con técnicas de optimización no lineal, el problema de detección de cáncer mamario a partir de imágenes de biopsias de células tumorales, logrando obtener una caracterización de la solución al problema de clasificación, y, por lo tanto, una posible automatización del diagnóstico.

0.2. Motivación

A nivel mundial en el año 2012 cerca de 521.000 mujeres murieron a causa del cáncer de mama[1]. En este sentido, el mundo médico ha sido claro al decir que la realización de exámenes como la mamografía de forma rutinaria, constituye un aspecto vital en la búsqueda de un diagnóstico oportuno. Esto, debido a que una proporción no menor de mujeres a lo largo del desarrollo de dicha enfermedad no presentan ningún tipo de síntoma.

Es necesario señalar que una mamografía es una imagen de la mama tomada con rayos X la cual es sometida a una evaluación visual donde los médicos buscan indicios de células cancerosas. En caso de hallar masas anómalas al interior de la mama se procede a realizar una biopsia, la que consiste en extraer una muestra del tejido, con la finalidad de evaluar si las células que lo conforman son benignas o malignas. Esto último suele estudiarse mediante el escrutinio visual bajo el microscopio del tejido extraído, así como también, a través de pruebas de laboratorio que por lo general toman días en generar resultados concluyentes, y, además, requieren de infraestructura médica sofisticada para llevarse a cabo.

En este contexto, se busca automatizar la clasificación de las células que componen el tejido obtenido en la biopsia, mediante el uso de una herramienta digital, práctica y efectiva que determine la naturaleza de las células que lo componen con la intención de reducir tiempos de espera y la alta demanda que dichos procedimientos sufren.

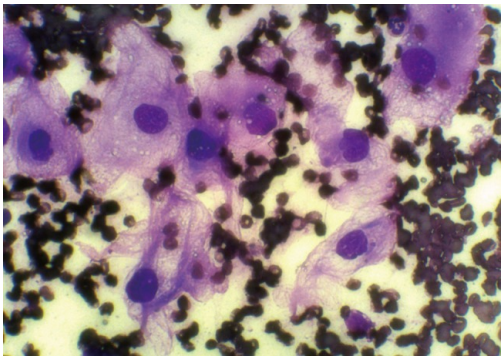

Figura 1: Biopsia con aguja fina de masa mamaria con presencia de células tumorales.

0.3. Formulación del problema de optimización

Se entenderá el problema de discernir cuando una célula es cancerígena o no, como un problema de clasificación de imágenes, sobre las cuales se establecerán métricas de interés biológicas (alta verosimilitud con la malignidad), como lo son, por ejemplo, el radio del núcleo celular (promedio de los segmentos radiales desde el centroide

[1] Global burden of cancer in womans

y su contorno), área (cantidad de píxeles al interior del contorno celular) o la textura del núcleo celular (varianza de la intensidad de la escala de grises de los píxeles que lo componen), entre otras, definidas en [7] que conformarán el espacio de características y la base de datos a utilizar, la cual se puede ver detenidamente en Bases de Datos.

Para dicha tarea, se hará uso del modelo SVM (support vector machines) introducido por Vapnik en los años 90 [6], que se basa en la idea de construir un hiperplano que divida los grupos de interés, y, cuya distancia con respecto a estos a su vez sea máxima.

Para entrenar nuestro modelo de clasificación, se debe tener un conjunto de ejemplos para que este logre aprender en base a ese conjunto. Se considerará $\{x_i, y_i\}_{i=1}^m$ un conjunto de datos, con $x_i \in \mathbb{R}^n$, compuesto por los atributos señalados con anterioridad (n=31, en nuestro caso, por lo que llamaremos al problema como uno de alta dimensionalidad), y, por otro lado, $y_i = \{+1, -1\}$, dependiendo de la clase binaria (benigno o maligno) a la cuál corresponde el vector x_i dadas sus características.

La idea es construir el hiperplano separador de clases, maximizando el margen que se genera entre los vectores de soporte[2] y el hiperplano, como se ilustra en la Figura 2.

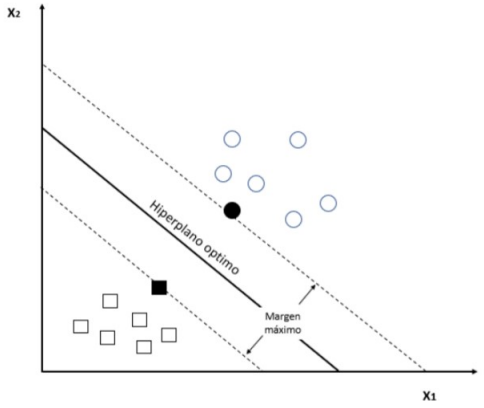

Figura 2: Hiperplano separador.

0.3.1. Hiperplano separador

Un hiperplano separador es una superficie de codimensión 1[3] de decisión que separa las distintas clases que tiene el problema de clasificación. El hiperplano es una ecuación de la forma:

$$w^T x + b = 0$$

0.3.2. Margen del hiperplano

El margen geométrico de un hiperplano $\mathcal{H}$ con respecto a la data, se define como la distancia más corta de un punto de entrenamiento x_i al hiperplano $\mathcal{H}$.

Observación: El mejor hiperplano $\mathcal{H}$, es el que tiene el mayor margen posible.

[2]Es el vector paralelo al hiperplano $\mathcal{H}$ tal que se encuentra justo en el margen o lo viola. Estos son los únicos que influyen sobre el hiperplano

[3]Es decir, es un subespacio de una dimensión menos que el espacio en el que se trabaja

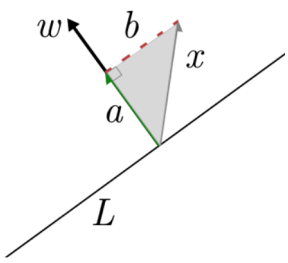

Figura 3: Construcción del margen.

Observación: Notemos que en la Figura 3, w juega el rol de vector normal del hiperplano. Además, a, es el resultado del producto interno entre x y w. Por otro lado, b es la distancia de x al hiperplano. De esta forma podemos dar paso a una formulación del problema de optimización.

0.3.3. Formulación

$$\begin{cases} \max_{w} \min_{x_i} \left| \langle x_i, \frac{w}{||w||} \rangle + b \right| \\ \textbf{s.t:} \quad sign(\langle x_i, w \rangle + b) = sign(y_i) \end{cases}$$

- Desventajas:
  1) Las restricciones no son lineales, ya que se comparan signos.
  2) Hay un máximo y un mínimo, por lo que se dificulta bastante el problema de optimización.
  3) La función objetivo es no lineal.

El objetivo será transformar este problema de optimización, en uno donde las restricciones sean desigualdades lineales y la función objetivo sea un polinomio cuadrático único.

0.3.4. Linealización de las restricciones

Para resolver este problema se utilizará la siguiente estrategia: Multiplicamos y_i en ambos lados de la expresión:

$$sign(< x_i, w > +b) = sign(y_i)$$

De esta forma existen 2 casos, el que nos interesa es cuando y_i tiene igual signo que $< x_i, w > +b$, entonces, se tendrá que:

$$(< x_i, w > +b)y_i \geq 0$$

y, así, el punto fue bien predicho por el Hiperplano H. De lo contrario, tendremos que:

$$(< x_i, w > +b)y_i \leq 0$$

Osea, el hiperplano habrá predicho de forma incorrecta la clase del punto x_i. De esta forma, la restricción es lineal, ya que y_i es una constante de entrada y las variables son los coeficientes de w.

Observación: Recordemos que el mejor Hiperplano es que el maximiza su margen, por lo que necesariamente este se debe encontrar justo en medio de los puntos más cercanos de cada clase.

0.3.5. Deshacerse del Max-Min

Notemos que w, tiene una libertad en cuanto a su escalamiento, es decir, podemos hacerlo tan grande o pequeño como queramos en su norma, sin que se vea afectado el hiperplano separador. Por lo que se puede hacer el siguiente artificio:

$$\langle x, w\rangle = ||w|| \cdot \langle x, \frac{w}{||w||}\rangle$$

Supongamos que tenemos el Hiperplano óptimo, con su vector normal w, no importa cuán cerca o lejos esté del punto de entrenamiento x, podriamos escalar $||w||$ de tal forma que $\langle x, w\rangle = 1$. Con esto podemos lograr que la distancia de un punto de entrenamiento x al hiperplano $\mathcal{H}$ esté dada por:

$$= ||w||\langle x, \frac{w}{||w||}\rangle = 1 \iff \frac{1}{||w||} = \langle x, \frac{w}{||w||}\rangle$$

De esta forma, si forzamos a que el punto más cercano tenga producto interno 1, todos los demás tendrán producto interno mayor o igual que 1. Una consecuencia de esto, es que la restricción cambia a ser $\langle x_i, w\rangle y_i \geq 1$ (ya no ≥ 0). Otra consecuencia, es que ya no nos interesa preguntarnos cuál es el punto más cercano a nuestro candidato a hiperplano óptimo, pues sabemos que está a distancia $\frac{1}{||w||}$. De hecho, si el punto óptimo no está a esa distancia, entonces el punto óptimo no cumple exactamente con la restricción (i.e, $\langle x, w\rangle > 1 \quad \forall x$), por lo que podríamos escalar w hasta que $\langle x, w\rangle = 1$, y, de esta forma habremos incrementado el margen en $\frac{1}{||w||}$.

En síntesis, el problema de optimización se reduce a maximizar $\frac{1}{||w||}$, o equivalentemente, a minimizar $||w||$.

Para evitar trabajar con la raíz cuadrada, lo haremos maximizando $||w||^2$, y multiplicando por un factor $\frac{1}{2}$ por conveniencia. Por lo que nuestro problema de optimización con la nueva función objetivo y la nueva restricción se expresa como:

$$\begin{cases} \quad min \quad \frac{1}{2}||w||^2 \\ \\ \textbf{s.t:} \quad (\langle x_i, w\rangle + b)y_i \geq 1 \end{cases} \tag{1}$$

De esta manera, a través de los cambios realizados se obtuvo un problema con restricciones de desigualdades lineales. Sin embargo, la función objetivo es una función cuadrática convexa, y, por tanto (SVM Duro) dado por (1) es un problema de programación cuadrática. Se puede señalar que una buena noticia, es que existen métodos generales para encontrar soluciones para estos problemas pero frecuentemente sufren problemas de estabilidad numérica y sus tiempos de ejecución son poco satisfactorios.

0.3.6. Desarrollo del problema

Dado el desarrollo del problema presentado anteriormente para SVM, notamos que la función del problema de optimización (SVM Duro) cumple con las condiciones de KKT, por lo que el problema de minimización equivale a resolver el problema de Multiplicadores de Lagrange. Una vez llegado a este punto, utilizando las restricciones del problema, es posible obtener la formulación dual para SVM (14), el cuál nuevamente es un problema de programación cuadrática, ampliamente estudiado en la literatura, por lo que se conocen muchos métodos para solucionar este tipo de problemas. Sin embargo, el poder encontrar un hiperplano que separe las clases, sólo es posible cuándo los datos son linealmente separables. Una pregunta natural es, ¿Qué hacemos si nuestro conjunto de datos no es linealmente separable?, para esto tendremos dos situaciones: La primera es cuando el conjunto es *casi linealmente separable*, es decir, cuando es posible separar los dos conjuntos de puntos por un hiperplano salvo por una cantidad pequeña de puntos. La segunda es cuando los datos bajo ningún punto de vista podrán ser separados por un hiperplano, a este caso, lo llamaremos *Problema no linealmente separable*.

Una solución para el problema cuando los datos son *casi linealmente separables*, consiste en reformular SVM agregando variables de holgura, las cuales permiten ignorar una pequeña fracción de los vectores de soporte, es decir, aquellos datos cercanos a la frontera de decisión, de forma que el hiperplano clasificador sea más robusto a la hora de adaptarse al patrón de los datos. Notemos que mientras más puntos de entrenamiento ignoramos, mayor será nuestro margen, es en este punto donde se genera un tradeoff entre la cantidad de puntos ignorados y el margen obtenido. Esto se puede controlar penalizando las variables de holgura en nuestra función objetivo, a través de la constente C en la ecuación (SVM Blando).

Dado lo anterior, a continuación el trabajo se enfocará en la formulación blanda de SVM, debido a la generalidad que presenta al extenderse a problemas que no son linealmente separables de manera estricta. La siguiente pregunta es, ¿Cómo encontrar la solución al problema de optimización de la formulación(SVM Blando)?, Notar que estamos trabajando en un espacio reflexivo donde tanto la función objetivo como la restricción son convexas y continuas, por tanto, gracias al teorema de KKT, el problema de minimización es equivalente a utilizar multiplicadores de Lagrange, luego, a través de las restricciones de esta formulación es que se puede obtener el problema dual para SVM Blando, el cual es un problema de programación cuadrática. Notar que la función a maximizar se puede escribir como:

$$\sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle = \frac{1}{2} \alpha^T \Lambda \alpha + q^T \alpha \tag{2}$$

Observación: La matriz $\Lambda \in \mathbb{R}^N \times \mathbb{R}^N$ es definida positiva, así, la función a maximizar es cóncava, por lo que se tiene que el máximo del problema es único.

La siguiente pregunta a abordar es, ¿Qué se puede hacer si el conjunto de datos no es casi linealmente separable en el espacio de características? Intuitivamente, podría suceder que pueda serlo en un espacio de mayor dimensionalidad, entonces, a través de una función (ϕ) podríamos aumentar la dimensionalidad del conjunto de datos, y, a continuación plantear el SVM. Sin embargo, el costo computacional asociado a esta técnica sería demasiado elevado. Esta es la motivación de introducir funciones que simulen elevar la dimensionalidad del problema, sin necesidad de transformar las propiedades originales de los datos. A este tipo de funciones, las denominaremos Kernel, las cuáles deben cumplir con lo siguiente:

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle \tag{3}$$

Dónde ϕ es una función que va de X a un espacio de mayor dimensionalidad dotado de un producto interno.

Notemos que, por lo general dada una transformación ϕ, definimos nuestro Kernel como en (3). Los esfuerzos

más recientes, proponen diseñar un Kernel, sin necesidad de fijar una función ϕ que transforme los datos. Esta idea, se basa en la intuición de que una transformación del espacio de características original debiese ser equivalente a cambiar la métrica con la cuál se comparan los datos originales, que en el caso del problema clásico viene dada por $\langle \cdot, \cdot \rangle$. Algunos de los ejemplos clásicos de Kernel se presentan en Kernels populares.

Finalmente, añadiremos a la formulación con holguras(blanda) la técnica del Kernel, para esto, es necesario notar que al observar el problema (26), los datos solamente se encuentran presentes en la función $\langle \cdot, \cdot \rangle$, por lo que en vez de transformar la dimensión de los datos y hacer SVM en el nuevo espacio, se puede considerar un Kernel que remplace el producto interno sobre el espacio original por un producto interno de una transformación no lineal de los datos.

En conclusión, el nuevo problema dual se escribe como la ecuación (27), el cual será nuestro problema a resolver. Para ello notamos que podemos reescribir el problema nuevamente como uno de programación cuadrática, osea de la forma:

$$\begin{aligned} \min \quad & \frac{1}{2}\alpha^T P \alpha + q^T \alpha \\ \text{s.t.} \quad & G\alpha \leq h, \\ & A\alpha = b \end{aligned} \tag{4}$$

La asignación explicita se define en el anexo Problema Cuadrático.

Al resolver el problema de programación cuadrática presentado anteriormente, se obtiene como resultado los multiplicadores de Lagrange α_i, por lo que la función de decisión para clasificar los datos viene dada por:

$$F(x) = sign(\sum_{i=1}(\alpha_i y_i \kappa(x_i, x)) + b^*) \tag{5}$$

Considerando, $B = \{\ell : 0 < \alpha_\ell < C\}$, (18) y (21) como $\xi_\ell = 0$, se obtiene:

$$\begin{aligned} b^* &= |B|^{-1} \sum_\ell \left(y^{(\ell)} - w^T x^{(\ell)} \right) \\ &= |B|^{-1} \sum_\ell \left(y^{(\ell)} - \sum_{\ell'} \alpha_{\ell'} y^{(\ell')} \left\langle x^{(\ell')} x^{(\ell)} \right\rangle \right) \end{aligned} \tag{6}$$

conocido como el bias o sesgo del modelo.

Observación: No todos los multiplicadores de Lagrange serán no nulos, por tanto diremos que aquellos x_i que estarán presente en la función de decisión (5) serán los vectores de soporte del problema.

0.4. Análisis al problema de optimización

A continuación, se presentarán los principales análisis al problema de optimización final (27).

0.4.1. Admisibilidad de Kernel

Intuitivamente la idea de subir la dimensión de los datos funciona, pues en el peor de los casos cada dato representará una dimensión y el problema será trivialmente separable. Esta idea la demostró Cover [1] a través del Teorema de Cover. El cual nos dice que la probabilidad de separabilidad lineal de los datos aumenta con probabilidad Bernoulli en la cantidad de datos. Sin embargo, redimensionar datos es un problema muy poco eficiente en memoria y la complejidad dependerá de la función utilizada.

Es por esto que el truco del Kernel es tan relevante, recordemos que un Kernel es admisible si es una función de los datos dependiente de una transformación en productor interno como en (3).

Ahora la interrogante es, cuáles funciones son admisibles como Kernel. Esto lo respondió Mercer en 1909 [3] en el contexto de la teoría de las ecuaciones integrales, estableciendo que un Kernel será admisible si y solamente si es positivo, es decir: Si es simétrico y si para cualquier n y cualquier $S = \{x^{(l),y^{(l}}\}_{l=1}^{n}$, su matriz Gram correspondientes es semi-definida positiva, es decir, si $\forall \alpha \in \mathbb{R}^n : \alpha^T K \alpha > 0$. Además, Mercer también demostró que dado un Kernel positivo es posible construir muchos mapas ϕ que satisfacen las propiedades requeridas, como coralario directo de su teorema, el cuál entrega una base ortonomal infinita del espacio con las cuáles construir otros Kernels.

Una corriente importante en la literatura de SVMs consiste en el aprendizaje del Kernel, idea muy en línea con lo que hoy se denomina metric learning. Este se motiva dado que la suma y multiplicación de Kernels es Kernel, y, por tanto existen dos formas principales de hacerlo, adtiva o multiplicativamente en base a Kernels positivos. Finalmente, del criterio de entrenamiento de una SVM según la función objetivo de (27) se puede aprender el Kernel según el siguiente algoritmo, teniendo en cuenta que $\mathcal{M} = \{\mu \geq 0 : \|\mu\|_p \leq \gamma\}$ y $p \geq 1$.

Algorithm 1: Algoritmo Genérico de Aprendizaje del Kernel.

1 Inicializar $\mu \in \mathcal{M}$.
2 **do**
3 Para μ fijo, actualizar α resolviendo

$$\max_{\alpha,\mu} W(\alpha, \mu) \text{ s.t. } \alpha \in \mathcal{C}.$$

4 Para α fijo, actualizar μ via gradiente $\mu \leftarrow \mu + \eta \frac{\partial W(\alpha,\mu)}{\partial \mu}$.
5 Proyectar μ en $\mathcal{M}$.
6 **while** *not convergence*;

0.4.2. Análisis de sensibilidad

Considerando el problema (26), notamos que para SVM de margen suave o blando, la existencia del hiperparámetro $C > 0$, nos surge la pregunta natural, ¿Cuánto podemos variar este parámetro?, y ¿cuáles son las consecuencias

de modificarlo?. Para responder a estas preguntas, haciendo uso del problema dual con la función Kernel asociada (27), vemos que el parámetro de interés ahora se encuentra en las restricciones del problema, es por esto que hacemos el estudio del análisis de sensibilidad para un problema general de programación matemática, cuando las restricciones son perturbadas. En este caso consideramos el problema de programación matemática sólo con restricciones de desigualdad. Así, llamaremos problema perturbado, al problema de progrmación matemática dónde se han aplicado perturbaciones en las restricciones. Bajo hipótesis de dualidad fuerte, tenemos la ventaja de que las variables duales óptimas nos pueden entregar valiosa información sobre la sensibilidad del problema al perturbar sus restricciones. Si Llamamos $p^*(u)$ al valor óptimo del problema perturbado, y consideramos x un punto factible para el problema perturbado, es decir, que cumple las restricciones del problema. Obtenemos la expresión (30):

De esto último se pueden desprender los siguientes análisis.

- Si tenemos que α_i^* es muy grande, y reforzamos la i-ésima restricción, es decir, $u_i < 0$, tendremos que el valor óptimo de $p^*(u)$ aumentará demasiado.
- En cambio, si α_i^* es pequeño y relajamos la i-ésima restricción, es decir, $u_i > 0$. Entonces $p^*(u)$ no disminuirá su valor en gran medida.

Finalmente, si consideramos el caso en el $p^*(u, v)$ óptimo del problema perturbado, es diferenciable en $u = 0$, bajo el supuesto de dualidad fuerte, es posible relacionar las variable óptima dual α^*, con el gradiente de $p^*(u)$ en $u = 0$. Tenemos que de (30), si tomamos $u = te_i = t(0, ..., 1, ..., 0)$ Obtenemos las expresión (31), que dan cuenta que los multiplicadores de Lagrange, son las sensibilidades del valor óptimo con respecto a las perturbaciones. Debemos notar que estas medidas son locales, es decir, válidas para una vecindad del óptimo de la función, y además, nos da nociones de que tan activas son algunas restricciones en el punto óptimo en cuestión.

Por ejemplo, si $g_i(x^*) > 0$ es una restricción no activa, podremos reforzarla o relajarla a conveniencia en pequeñas cantidades, sin que cambie el valor óptimo del problema. Es aquí donde la holgura complementaria, nos dirá que el multiplicador de Lagrange asociado es 0. Por otro lado, si $g_i(x^*) = 0$ entonces la restricción es activa en el óptimo, y, en consecuencia el multiplicador de Lagrange será no nulo. Aplicando esto a nuestro problema, dónde es de nuestro interés mover el parámetro $C > 0$ de la restricción $C \geq \alpha_i > 0$, notamos que, para un valor de C muy grande, el multiplicador de Lagrange asociado tendrá demasiada libertad, esto hará que si la restricción es activa, es posible que el óptimo de la función cambie demasiado dado un alto valor del multiplicador α_i asociado a la i-ésima restricción. Cabe destacar que si $C \to \infty$, se tiene el problema de SVM de margen duro, el cuál es muy sensible a los datos, lo que se refleja en este análisis de sensibilidad, dado que el multiplicador asociado puede cambiar drásticamente el valor de la función objetivo. Contrariamente, si el valor de C es muy cercano a cero, α_i asociado verá reducido su nivel de libertad, por lo que no se verá afectado el óptimo de la función objetivo.

Debemos notar, que el rol de esta constante $C > 0$, es penalizar los errores determinando el margen, por lo que un valor de C muy grande hará que el algoritmo intente no equivocarse al clasificar, por lo que se espera que la cantidad de support vectors, sea pequeña. En cambio, si el valor de la penalización es muy pequeña, no será costoso para el algoritmo realizas una mala clasificación, por lo que eventualmente, se espera que el algoritmo reconozca todos los puntos como support vectors.

Como ya se mencionó anteriormente, el rol que cumple la constante C es penalizar los errores que comete el algoritmo al momento de clasificar, geométricamente lo que hace esta penalización es expandir o contraer el margen de nuestro hiperplano, dependiendo del valor escogido para C, el poder controlar el margen a través de un hiperparámetro es de gran interés, pues recordemos que mientras más grande sea el margen, más robusto será nuestro clasificador. Al expandir el margen del hiperplano, tenemos que más imágenes

serán consideradas como support vectors, o imágenes confusas al momento de predecir, ya que estarán en el interior o justo en la frontera del margen. Pero, ¿Cómo ayuda al algoritmo tener más imágenes similares entre clases?. La respuesta es que al tener más ejemplos difíciles de clasificar, el algoritmo puede discernir mejor cuándo se le presenten nuevos ejemplos de una clase pero con características muy similares a las de su clase contraria.

0.4.3. Aproximación de support vectors

Como vimos anteriormente, la función de decisión (5) resulta ser la clasificación que asigna el hiperplano separador obtenido. Notamos entonces que en este, los únicos vectores presentes son aquellos que tienen multiplicadores de Lagrange no nulos, en otra palabras, el problema de clasificación se reduce a obtener aquellos vectores llamados usualmente de soporte. Como solamente se necesita un pequeño subconjunto de los datos, que en el contexto de esta investigación, serían algunas de las imágenes de biopsia, para lograr establecer un clasificador de cáncer mamario. Surge entonces la interrogante, si existirá algún algoritmo para determinar los support vectors sin tener que lidiar con el problema de optimización.

La intuición detrás de los support vectors se basa en el hecho de determinar los puntos más cercanos de una clase con respecto a la otra, por tanto los vectores candidatos de una clase serán aquellos más similares a la otra clase. En nuestro contexto entonces notamos que las imágenes que son candidatas a ser support vector serán por ejemplo, el conjunto de biopsias malignas que tienen características similares a las benignas y viceversa. Por supuesto, asumiendo que la ingeniería de atributos elaborada por los médicos es lo suficientemente correcta. Obtener aquellas imágenes, puede ser un trabajo muy difícil pero de gran utilidad, pues obtener este tipo de biopsias nos permitiría tener un algoritmo más robusto y de paso automatizar el diagnostico de cáncer mamario.

0.5. Rutinas de resolución

0.5.1. Resolución como problema de Programación Cuadrática

Teniendo en cuenta la formulación dual de nuestro problema inicial de Optimización podemos ver que es posible resolverlo mediante métodos tradicionales:

```

Se definen las matrices y vectores del problema cuadr tico (2.3) asociados al problema
(3.21) como se explicita en anexo 3.08

def linear_Kernel(x, z):
    return x*z.T

def fit(X, y, Kernel, C):
    cvxopt.solvers.qp(P, q, G, h, A, b) <- resuelve el problema (3.21)
    a <- multiplicadores de lagrange no nulos definen los vectores de soporte
    return a

Se establece un criterio de convergencia para fit(X, y, Kernel, C)

b<- se calcula como en (2.5)
predicciones <- signo de la funci n decisi n (2.4)
```

Listing 1: QP

Puede revisarse el código (Problema cuádratico) en detallen en: Repositorio del proyecto

0.5.2. Resolución mediante SVM

Para este método se procede mediante la librería Scikit Learn tradicional para aplicaciones de métodos de machine learning, para poder medir el rendimiento del modelo se procede a dividir el conjunto de imágenes de Biopsia en entrenamiento y testeo.

```
1
2  Se definen los conjuntos de entrenamiento y testeo a partir del data set disponible.
3  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size, random_state)
4  Se eleige un Kernel y una constante de regularizaci n del margen blando.
5  svm = SVC(Kernel, C).fit(X_train,y_train) #entrenamiento.
6  #Grafica de support vectors proyectados a dos dimensiones.
7  ax.scatter(svm.support_vectors_)
8  y_pred_test = svm.predict(X_test) #prediccion
9  #Visualizacion de Resultados
10 plot_confusion_matrix(svm, X_test)
```

Listing 2: SVM

Puede revisarse el código (SVM) en detallen en: Repositorio del proyecto

0.5.3. Extensiones

Cómo se mencionó anteriormente, los problemas de programación cuadrática han sido ampliamente estudiados en la literatura, contando con una sólida base teórica que garantiza la existencia y unicidad de soluciones bajo ciertas condiciones. Así, también, se cuenta con múltiples herramientas computacionales que permiten resolver este tipo de problemas, como por ejemplo el Solver de programación cuadrática de CVXOPT en Phyton, el cuál emplearemos en este trabajo.

Dentro de las muchas estrategias para resolver el problema, se encuentra, el método de descenso de gradiente sobre la función de pérdida u otras variantes de métodos de descenso, tales como, gradiente conjugado o gradiente estocástico.

Dada la buena performance del SVM en su formulación dual del problema con holgura y Kernelizada (27), se dice que es el estado del arte en problemas de clasificación. Sin embargo, se ve limitado por su velocidad de entrenamiento para conjuntos de datos masivos. Dado que el hiperplano construido solo depende de los vectores de soporte, es decir, aquellos datos que están cerca de la frontera de decisión,debido a esto es que se han propuesto algoritmos mas eficientes [9] que trabajan quitando aquellos datos que no tendrían efectos sobre la función de decisión, manteniendo las condiciones de KKT y teniendo como ventajas la reducción de la complejidad computacional y la aceleración de convergencia del algoritmo.

Por otro lado, se tiene que SVM con programación lineal resulta ser muy efectivo cuando se trata de grandes cantidades de datos, a diferencia de su formulación como problema de programación cuadrática, sin embargo, no se sabe mucho acerca de la convergencia del problema formulado con programación lineal. En el trabajo [4], se presentan resultados acerca del comportamiento de convergencia de la SVM de programación lineal, mostrando que es casi el mismo que el de la SVM de programación cuadrática, presentando además un límite superior para el error de clasificación errónea de datos. En esta misma linea, es que se formulan variantes del problema de SVM, como, por ejemplo, Least Squares SVM (LS-SVM) [5], dónde se propone utilizar una función de costo de mínimos cuadrados para obtener un conjunto lineal de ecuaciones en el espacio dual. Al comparar SVM con LS-SVM [8], se obtiene como resultado que bajo ciertas condiciones, LS-SVM para clasificaciones de clases binarias es equivalente al SVM de margen duro basado en la conocida métrica de Mahalanobis, más aún, utilizando herramientas de teoría asintótica de distribuciones, sobre los valores propios de la matriz de covarianza de los datos, es posible demostrar que LS-SVM es equivalente a SVM con la métrica euclidiana típica.

Además recientemente, dado que (27) presenta las dificultades anteriormente mencionadas es que se logró [2] mejorar más aún la convergencia del método. Esto, se obtuvo linealizando el problema por medio de aproximaciones empíricas del Kernel, calculadas de forma eficiente a través de descomposiciones de rango bajo, es decir, aproximando la matriz de Kernel por medio de una descomposición en vectores propios que minimice el rango.
Por otro lado, puesto que (27) es un problema de optimización cóncava, una variante clásica para resolver el problema es el algoritmo de Frank-Wolfe, el cual es ampliamente ocupado en el caso de SVM no lineal y de gran escala (lo cual sería de gran ayuda en caso de que se lograran automatizar los análisis de biopsias). Así, recientemente se han propuesto nuevos algoritmos basados en FW con ventajas teóricas como las garantías en términos de convergencia y número de iteraciones, así como también ratios de convergencia lineales [10].

0.6. Representación y caracterización de soluciones

Con la finalidad de evidenciar cómo se pierde la noción geométrica del problema cuando la dimensión de este aumenta a más de 3, a través del método de *Análisis de componentes principales (PCA)*, dónde se escogen dos componentes principales, con el fin de hacer una proyección de los atributos en un plano para así poder visualizar cómo se comportan los support vectors, es decir, poder identificar cuáles son las imágenes de mamografías que definirán la regla para clasificar nuevas imágenes.

En la siguiente imágenes se presenta la proyección de todos los atributos de cada imagen en sus 2 componentes principales, variando el parámetro C del problema de SVM de margen suave, desde valores grandes, hasta valores muy pequeños, dónde cada clase es identificada con un color y se distinguen los support vectors con un anillo alrededor del punto.

Tal como se menciona en el análisis de sensibilidad del problema, para valores grandes de C se espera tener una menor cantidad de support vectors, mientras que para valores de C muy pequeños se espera que todos las imágenes sean consideradas support vectors por el algoritmo.

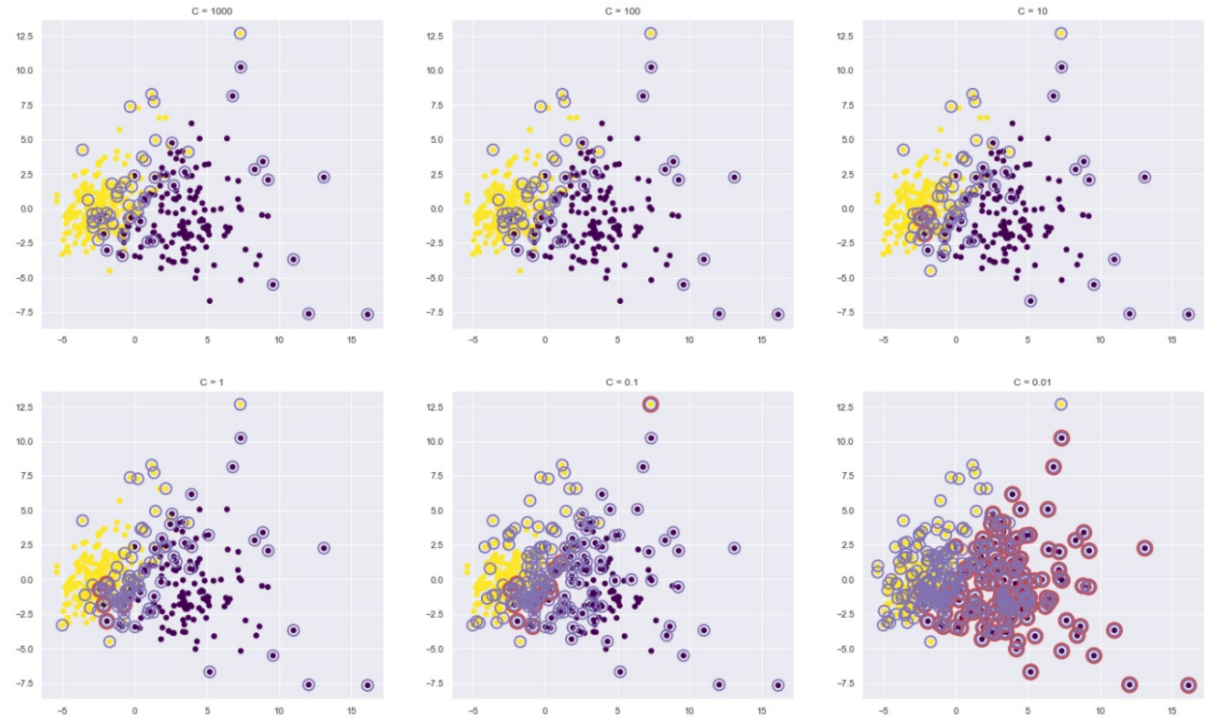

Figura 4: Identificación de vectores de soporte para distintas contantes de regularización.

Notemos que para los casos dónde C tiene un valor adecuado, ya no es posible caracterizar visualmente a los support vectors como los vectores más cercanos a la frontera de decisión, ya que si bien cumplen esto en dimensiones mayores o iguales a la dimensionalidad del problema, al considerar la proyección en el plano realizada por PCA, se pierde esta noción geométrica.

Otro punto importante, es la introducción de un tipo de Kernel, recordemos que esto se hace con la finalidad

de combatir el problema de la maldición de la dimensionalidad. Para efectos de este trabajo, se ha realizado la construcción del hiperplano utilizando dos tipos de Kernel, gaussiano ('rbf'), y lineal ('linear'), tanto para el problema cuadrático como para SVM de sklearn. Para este último, se obtuvo como resultado que si bien sus predicciones correctas, tanto de casos malignos como benignos son bastante similares, nos encontramos con una ventaja al usar Kernel gaussiano, ya que este no tiene predicciones de casos falsos negativos, a diferencia de Kernel lineal, que cuenta con 5 casos falsos negativos. En el contexto de nuestro problema de predicción del cáncer, un falso negativo provoca que un paciente con células cancerígenas malignas sea diagnosticado como un paciente con células cancerígenas benignas, pudiendo tener consecuencias fatales, es por esto que se busca minimizar la cantidad de falsos negativos. Por otro lado, para el problema cuadrático, tenemos kernel lineal, que en comparación con SVM (kernel lineal) tiene un mayor error al predecir casos benignos prediciendo maligno, sin embargo, comete menos equivocaciones en los falsos negativos, que recordamos, es el caso que buscamos minimizar. Mientras que para kernel gaussiano, su comportamiento es considerablemente peor con respecto al de SVM, contando con 4 casos falsos negativos, los cuáles son altamente penalizados para efectos del diagnostico del cáncer.
Todos los resultados de las matrices de confusión pueden encontrarse en el Repositorio del proyecto

0.7. Conclusión

En este trabajo se ha abordado el problema de SVM desde su construcción con una intuición geométrica, pasando por sus formulaciones como problema de optimización y sus posibles variantes con la introducción de herramientas matemáticas más sofisticadas, como el uso de diferentes tipos de Kernel en reemplazo del producto interno, y penalizaciones con el fin de relajar el problema.

Es en el apartado de la formulación de SVM como un problema de programación cuadrática donde se identifica la principal falencia del método, y, es que para grandes bases de datos, SVM sufre de un tiempo de entrenamiento demasiado elevado, debido a que el algoritmo debe encontrar los multiplicadores de Lagrange asociados a cada imagen para poder construir el hiperplano que mejor separa ambas clases, además de que debe almacenar todos los cálculos del Kernel en memoria ($\mathcal{O}(n^2)$, con n la cantidad de datos). Por otro lado el solver qp usa métodos de punto interior lo cuál presenta una convergencia lenta, en complejidad computacional $\mathcal{O}(n^3)$. Esta situación la comprobamos generando data artificial en la implementación del problema cuadrático disponible en el [4], suponiendo que en vista al futuro se incorporarán cada vez más imágenes al modelo, testeamos con 300,000 imágenes, sin lograr el entrenamiento por falta de memoria. Luego de ello se comprobó con 5,000 nuevas imágenes, demorando aproximadamente 13 minutos su entrenamiento.

Sin embargo, se evidenció que los únicos vectores que realmente son relevantes en la construcción de este hiperplano, son los llamados support vectors, imágenes de mamografias que tienen multiplicador de Lagrange asociado, no nulo, es decir, el SVM se está entrenando con muchos ejemplos que no aportan información al momento de la construcción del hiperplano como función de clasificación. Por lo que la pregunta que surge naturalmente es, ¿Existe alguna forma de poder encontrar cuales son las imágenes relevantes?, o en su defecto, ¿Es posible aproximar estos multiplicadores?. Ya que de esta forma, podríamos reducir sustancialmente la data con la que se entrena SVM, agilizando el proceso ante una posible automatización del diagnóstico.

Finalmente, siguiendo la línea de la aproximación de multiplicadores de Lagrange, una interesante propuesta es darle un enfoque probabilístico, dónde la idea es encontrar un estimador máximo verosímil para los multiplicadores de Lagrange, el cuál dependa únicamente de los datos. Una posible investigación que se propone, es aprender la distribución que siguen los multiplicadores de Lagrange, asumiendo alguna distribución a priori, y así, a través de la distribución de los datos y aplicando la regla de Bayes, aprender la distribución a posteriori de los multiplicadores de Lagrange asociado a cada imagen. Una vez obtenida la distribución, obtener algún estimador máximo verosímil, o en su defecto, algún estimador con buenas propiedades para la reducción de data.

[4] Repositorio

Bibliografía

[1] T. Cover. *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition*. 1965.

[2] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang, and K. Zhang. Scaling up kernel svm on limited resources: A low-rank linearization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 30(2):369–378, 2019.

[3] James Mercer. *Functions of positive and negative type and their connection with the theory of integral equations*. 1909.

[4] Ding-Xuan Zhou Qiang Wu. *SVM Soft Margin Classifiers: Linear Programming versus Quadratic Programming*. Massachusetts Institute of Technology, 2005.

[5] T. VAN GESTEL, J. SUYKENS, B. BAESENS, S. VIAENE, J. VANTHIENEN, G. DEDENE, B. MOOR, and J. VANDEWALLE. Benchmarking least squares support vector machine classifiers. *Kluwer Academic Publishers*, 2004.

[6] Vladimir Vapnik. *Statistical Learning Theory*, volume 10. 1998.

[7] W.H. Wolberg W.N. Street and O.L. Mangasarian. *Nuclear feature extraction for breast tumor diagnosis*. International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

[8] J. Ye and T. Xiong. Svm versus least squares svm. *Proceedings of Machine Learning Research*, 2(2):644–651, 2007.

[9] Ying Zhang, Xizhao Wang, and Junhai Zhai. *A Fast Support Vector Machine Classification Algorithm Based on Karush-Kuhn-Tucker Conditions*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[10] Ricardo Ñanculef, Emanuele Frandi, Claudio Sartori, and Héctor Allende. A novel frank–wolfe algorithm. analysis and applications to large-scale svm training. *Information Sciences*, 285:66 – 99, 2014. Processing and Mining Complex Data Streams.

0.8. Anexo

0.8.1. Datos Linealmente Separables

Decimos que un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$, con clases binarias, es linealmente separable si existen $\gamma > 0$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ tales que $\forall i = 1, ..., n$

$$(\langle x_i, w \rangle + b) y_i \geq \gamma > 0 \tag{7}$$

0.8.2. El Langrangiano y condiciones de KKT SVM duro

El operador Lagrangiano viene dado por

$$\min_{w} \max_{\alpha_i \geq 0} L(w, b, \alpha) = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} (\alpha_i(\langle x_i, w\rangle + b)y_i - 1)) \tag{8}$$

Y planteando las restricciones tenemos planteado el siguiente problema:

$$\frac{\partial L(w, b, \alpha)}{\partial w} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \alpha_i x_i y_i = w \tag{9}$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{10}$$

$$\alpha_i((\langle x_i, w\rangle + b)y_i - 1) = 0 \tag{11}$$

$$\alpha_i \geq 0 \tag{12}$$

$$(\langle x_i, w\rangle + b)y_i \geq 1 \tag{13}$$

Donde tenemos que (3.5) corresponde a la restricción de *holgura complementaria*, (3.6) corresponde a condiciones de multiplicadores, y (3.7) es la restricción de nuestro problema original o factibilidad.

0.8.3. Problema Dual Duro

Utilizando la restricciones (3.4) y (3.5) en la Lagrangiana establecemos la formulación dual:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i = 1, 2, ..., n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \forall i = 1, ..., n \end{aligned} \tag{14}$$

De esta manera podemos reformular nuetro problema reemplazando el producto internos por nuestra funcion Kernel quedando de la forma:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i = 1, 2, ..., n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \forall i = 1, ..., n \end{aligned} \tag{15}$$

0.8.4. SVM blando

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & (\langle x_i, w\rangle + b)y_i \geq 1 - \xi_i, \\ & \xi_i \geq 0 \forall i = 1, ..., n \end{aligned} \tag{16}$$

0.8.5. El Langrangiano y condiciones de KKT SVM Blando

En términos de multiplicadores de Lagrange:

$$\min_{w} \max_{\alpha_i \geq 0, \mu_i \geq 0} L(w, b, \alpha, \mu) = \min_{w} \max_{\alpha_i \geq 0, \mu_i \geq 0} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n}(\alpha_i(((\langle x_i, w\rangle + b)y_i - 1 + \xi_i)) - \sum_{i=1}^{n} \mu_i \xi_i \tag{17}$$

y las condiciones de KKT:

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \alpha_i x_i y_i = w \tag{18}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{19}$$

$$\frac{\partial L(w,b,\alpha)}{\partial \xi} = 0 \quad \Longrightarrow \quad C - \alpha_i - \mu_i = 0 \tag{20}$$

$$\alpha_i(((\langle x_i, w\rangle + b)y_i - 1 + \xi_i) = 0 \tag{21}$$

$$\alpha_i, \mu_i \geq 0 \tag{22}$$

$$(\langle x_i, w\rangle + b)y_i \geq 1 - \xi_i \tag{23}$$

$$\mu_i \xi_i = 0 \tag{24}$$

$$\xi_i \geq 0 \tag{25}$$

0.8.6. Problema Dual Blando

Reemplazando las restricciones (3.12),(3.13) y (3.14) en la Lagrangiana obtenemos la formulación dual:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_{i}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \quad \forall i = 1, 2, ..., n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \forall i = 1, ..., n \end{aligned} \tag{26}$$

Reformulando con la función Kernel:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_{i}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \quad \forall i = 1, 2, ..., n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \forall i = 1, ..., n \end{aligned} \tag{27}$$

0.8.7. Kernels Populares

- Lineal: $k(x, x') = x^T x'$
- RBF o Gaussiano de escala γ: $k(x, x') = e^{-\gamma \|x^T x'\|^2}$
- Polinomial de grado p, intercpto c_0 y escala γ: $k(x, x') = (\gamma x^T x' + c_0)^p$
- Laplaciano de escala γ: $k(x, x') = e^{-\gamma \|x^T x\|_1}$

0.8.8. Problema cuadrático

Se reescribe el problema dual de SVM blanco con Kernel hyperref[eq: dual blando](27) como un problema de programación cuadrática (4):

- $$P = yy^T K = \begin{pmatrix} y_1y_1 & \dots & y_1y_n \\ y_2y_1 & \dots & y_2y_n \\ \vdots & \ddots & \vdots \\ y_ny_1 & \dots & y_ny_n \end{pmatrix} \begin{pmatrix} \kappa(x_1,x_1) & \dots & \kappa(x_1,x_n) \\ \kappa(x_2,x_1) & \dots & \kappa(x_2,x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n,x_1) & \dots & \kappa(x_n,x_n) \end{pmatrix}$$
- $$q = [-1]_{1\times n}$$
- $$G = \begin{pmatrix} -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} -I_n \\ I_n \end{pmatrix}_{2n\times n}$$
- $$h = \begin{pmatrix} [0]_{1\times n} & [C]_{1\times n} \end{pmatrix}_{1\times 2n}$$
- $$A = y$$

$$b = 0$$

0.8.9. Problema de programación matemática

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \qquad i = 1,\dots,m \end{aligned} \tag{28}$$

0.8.10. Problema de programación matemática perturbado

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq u_i \qquad i = 1,\dots,m \end{aligned} \tag{29}$$

0.8.11. Condición de optimalidad del problema perturbado

$$\begin{aligned} p^*(0) = \inf_x \mathcal{L}(x, \alpha^*) &\leq f(x) + \sum_{i=1}^{m} \alpha_i^* g_i(x) \\ &\leq f(x) + \sum_{i=1}^{m} \alpha_i^* u_i \\ &= f(x) + \alpha^{*T} u \end{aligned}$$

Obteniendo que

$$p^*(u) \geq p^*(0) - \alpha^{*T} u \tag{30}$$

0.8.12. Sensibilidad para el problema perturbado

- para $t > 0$

$$p^*(te_i) - p^*(0) \geq -t\lambda_i^* \iff \frac{p^*(te_i) - p^*(0)}{t} \geq -\alpha_i^*$$
$$(t \searrow 0) \Longrightarrow \frac{\partial p^*}{\partial u_i}(0) \geq -\alpha_i^*$$

- para $t < 0$

$$p^*(te_i) - p^*(0) \geq -t\alpha_i^* \iff \frac{p^*(te_i) - p^*(0)}{t} \leq -\alpha_i^*$$
$$(t \nearrow 0) \Longrightarrow \frac{\partial p^*}{\partial u_i}(0) \leq -\alpha_i^*$$

Concluyendo de esta forma que

$$\frac{\partial p^*}{\partial u_i}(0) = -\alpha_i^* \tag{31}$$

0.8.13. Teorema de Cover

Sea ϕ una función de $\mathbb{R}^d$ a $\mathbb{R}^D$, $M \in \mathbb{Z}$. Para cualquier conjunto de $n \in \mathbb{Z}$, ejemplos $S = \{(x^{(l)}, y^{(l)})\}_{l=1}^n$, $x^{(l)} \in \mathbb{R}^d$, $y^{(l)} \in \{\pm 1\}$, tenemos que $\phi(S) = \{(\phi(x^{(l)}), y^{(l)})\}$ es linealmente separable con probabilidad:

$$P(n, D) = \frac{1}{2^{n-1}} \sum_{k=0}^{D-1} \binom{n-1}{k}$$

0.8.14. Teorema de Mercer

Para cualquier Kernel positivo $k : \mathbb{X} \times \mathbb{X}$, existe una colección (posiblemente infinita) de funciones ortonormales $\{\psi_k\}_k \, \psi_k \in L_2(\mathbb{X})$ y constantes $\{\lambda_k\}_k \, , \lambda_k > 0$, tales que

$$k\left(x^{(i)}, x^{(j)}\right) = \sum_k \lambda_k \psi_k\left(x^{(i)}\right) \psi_k\left(x^{(j)}\right)$$

para (casi) todo $x^{(i)}, x^{(j)} \in \mathbb{X}$. Además, las funciones $\{\psi_k\}_k$ corresponden a las funciones propias del operador integral de Hilbert-Schmidt $T_k : L_2(\mathbb{X}) \to L_2(\mathbb{X})$

$$T_k[f](x) = \int_{\mathbb{X}} k\left(x, x'\right) f(x) f\left(x'\right) d\mu\left(x'\right)$$

y las constantes $\{\lambda_k\}_k$ son sus correspondientes valores propios.

0.8.15. Bases de Datos

Actualmente contamos con dos bases de datos que evalúan diversas métricas pertinentes con el diagnóstico de cáncer mamario[5]:

- **Primera base de datos**: Esta base de datos de cáncer mamario fue obtenida desde el Hospital de la Universidad de Wisconsin y consta de 699 observaciones sobre las cuales se consideran 12 atributos. De los cuales uno de ellos es un número de identificación o ID del examen, otro atributo presente es la etiqueta del resultado del estudio de la célula (Benigna-Maligna) y el resto de ellos son métricas que describen características morfológicas de estas y que se concederán de interés para este estudio. Los valores de dichas métricas constan de un número entero comprendido entre 1 y 10.

	code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2

Figura 5: Muestra de la primera base de datos.

- **Segunda base de datos**: Esta segunda base de datos proviene del Centro de Ciencias Clínicas de la Universidad de Wisconsin y se origina a partir del estudio de una imagen digitalizada de los núcleos celulares presentes en el tejido mamario extraído mediante el uso de una aguja fina. Esta data consta de un total de 569 observaciones sobre las cuales se consideran 32 atributos, como por ejemplo, el número identificador del examen realizado o ID del examen, una etiqueta con el resultado del estudio (Benigna-Maligna) y el resto consta de 30 métricas que describen distintas características del núcleo celular observado y que se consideran de interés para este estudio y cuyos rangos varían dependiendo de la métrica que se desee estudiar [7].

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980

Figura 6: Muestra de la segunda base de datos.

[5] Repositorio del proyecto con los datos