AA-Cuestionario-2.pdf

Anónimo

Aprendizaje Automático (Especialidad Computación y Sistemas Inteligentes)

3º Grado en Ingeniería Informática

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Cuestionario 2

1- Trabaja para una empresa que explota un criadero artificial de diversas especies de peces en un lago. La empresa desea construir un modelo para predecir los kilos de pescado comercializables en tres instantes de tiempo futuro, a una semana, a un mes y a tres meses. Para ello usará datos de muestras extraídas del lago. Las muestras las obtienen usando información disponible sobre la distribución de los bancos de peces en el lago a lo largo de las horas del día y usando los aparejos de pesca que se usan para explotación comercial del criadero. Analize la situación, haga las hipótesis que considere necesarias y diga que probabilidad hay de que la empresa obtenga los resultados que desea.

No es posible.
Independientemente de cualquier consideración práctica no contempla con detalle en el enunciado, el propio enunciado si contempla condiciones inadmisibles que impiden una correcta predicción en los periodos considerados.
La toma de muestra con las redes usadas para la comercialización introducen necesariamente un sesgo en los datos en aquellos tamaños que no son capturados por dichas redes. Obviamente los más pequeños que no son los necesarios para predecir a largo tiempo.

2- Considere un problema de clasificación. Identifique las condiciones bajo las cuales elegiría usar cada una de las siguientes aproximaciones (a,b,c) para resolverlo. Responda a (d):

a) El problema primal SVM-Hard
b) El problema dual SVM-Hard
c) El problema dual de SVM-Hard con uso de kernel
d) ¿Qué condiciones del problema de clasificación le harían desistir del uso del algoritmo SVM-Hard?

Nunca se sabe nada sobre la separabilidad de los datos. No es un argumento que tenga utilidad en este contexto.
1.- Problemas con datos muestrales de menor dimensionalidad que el tamaño de la muestra. La optimización es en el espacio de las muestras, con eficiencia O(N^3)

2.- Problema con datos muestrales de mayor dimensionalidad que el tamaño de la muestra. La optimización es en el espacio de los multiplicadores, con eficiencia O(N^3)
3.- Cuando queremos resolver la optimización en un espacio de alta dimensionalidad sin tener que añadir a mano funciones no lineales de las muestras.
4.- Dada la eficiencia de dicho algoritmo, O(N^3), el número de muestras es el factor más relevante.

3- Identifique las ventajas e inconvenientes del algoritmo SVM-soft, respecto de los otros clasificadores binarios que hemos estudiado: perceptron, regresión logística y redes neuronales. Analice el caso separable y no separable.

La estrecha relación entre el algoritmo SVM-soft y los algoritmos Perceptron (P), RL y RRNN se establece con la formulación de SVM-soft como un problema de regularización. En ese contexto, el análisis comparativo se centra en la función de pérdida usada por cada uno de ellos, ya que el término de regularización es común a todos ellos (minimizar la norma del vector de pesos).
Caso separable: En este caso SVM-soft alcanza la solución óptima en términos de menor dimensión de VX, ya que su función de perdida busca el hiperplano solución con máxima anchura entre clases. Las funciones de pérdida de P y RRNN solo buscan una solución separable. Por otro lado RL optimiza una pérdida extraída de la optimización de la verosimilitud de la muestra que no implica la solución de mayor separabilidad entre las distribuciones de cada clase.
Caso no separable: SVM-soft nos permite a través de la constante C del término de penalización de errores, elegir un compromiso entre error y generalización. Para cada valor de C la solución de SVM-soft es óptima ya que maximiza la separabilidad entre clases. Los demás algoritmos no disponen de dicho mecanismo de compromiso y por tanto sus soluciones no tienen ninguna garantía de optimalidad o de encontrar una solución aceptable. Sin embargo, RL ha mostrado experimentalmente que en los casos de alto nivel ruido la optimización de la verosimilitud de la muestra conduce a soluciones mejor regularizadas que la optimización por separabilidad.

4- Consideremos dos transformaciones no lineales de dimensión finita Φ_1 y Φ_2 y sus correspondientes núcleos K_1 y K_2. Definimos dos nuevas transformaciones como $\Phi_3(x) = (\Phi_1(x),\Phi_2(x))$ y $\Phi_4 = \Phi_1(x) \cdot \Phi_2(x)^T$.

a) Expresar los kernels correspondientes a Φ_3 y Φ_4 en términos de K_1 y K_2.
b) ¿Qué conclusiones saca respecto de $K_1 + K_2$ y $K_1 \cdot K_2$?

1.- $\Phi_3(x) = (\Phi_1(x),\Phi_2(x))$ por tanto $K_3(x,x') = (\Phi_3(x))^T\Phi_3(x') = (\Phi_1(x),\Phi_2(x))^T [^{\Phi1(x')}_{\Phi2(x'))}] = K_1(x,x') + K_2(x,x')$
2.- $\Phi_4(x) = \Phi_1(x)\Phi_2(x)^T$ por tanto $K_4(x,x') = (\Phi_4(x))^T\Phi_4(x') = (\Phi_1(x)\Phi_2(x')^T = \Phi_2(x)\Phi_1(x)^T\Phi_1(x')\Phi_2(x')^T = \Phi_2(x)K_1(x,x')\Phi_2(x')^T = K_1(x,x')K_2(x,x')$
3.- Dado que las matrices de Gram (covarianza) calculadas a partir de "kernels" son semidefinida positiva y simétricas, la suma y el producto permanecen semidefinidas positivas y simétricas. Por tanto si K_1 y K_2 son "kernels" también lo son K_3 y K_4 .

5- Discuta el siguiente razonamiento y posicionese a favor o en contra del mismo identificando los problemas o los soportes de la teoría para su aceptación o rechazo.
"Considere la tarea de seleccionar una regla de vecino más cercano. Para ello definimos el conjunto de hipótesis H_{nn} con N hipótesis, las k-NN reglas que se usan para $k=1,...,N$. Usamos el error en la muestra de cada regla para elegir el valor de k que minimiza E_{in} . Entonces usando la cota de E_{out} ya estudiada para el caso de un conjunto finito de hipótesis podemos concluir que $E_{in} \rightarrow E_{out}$ porque $logN/N \rightarrow 0$. Por lo tanto, concluimos que asintóticamente con N, estaremos eligiendo el mejor valor de k, basado solo en E_{in} ".

Supongamos una muestra de datos de tamaño N. Analicemos la conducta de la cota para este caso.
La cota para clases finitas es $E_{out} \leq E_{in} + O(\sqrt{(\frac{logM}{N}))}$ ahora $M = N$ ya que en cada caso se usan todas las k_NN reglas.
Para N asintótico $O(\sqrt{(\frac{logM}{N}))} \rightarrow 0$ pero dado que H_{nn} contiene siempre la regla 1-NN con $E_{in} = 0$ siempre será elegida esta regla para todo valor de N.

Pero sabemos que 1-NN es una regla que sobreajusta los datos por tanto no podemos decir que estamos eligiendo la mejor regla asintóticamente con ***N***. El resultado que garantiza $E_{in}(g) \rightarrow E_{out}(g)$ obliga a que $k(N) \rightarrow \infty$ y en este supuesto no es el caso.

6- Considere un modelo de red neuronal con dos capas totalmente conectadas: ***d*** unidades de entrada, n_H unidades ocultas y ***c*** unidades de salida. Considere la función de error cuadrática definida por

$J(w) \equiv \sum_{k=1}^{c} (t_k - s_k)^2 = \frac{1}{2}||t - s||^2$, donde el vector ***t*** representa los valores de la etiqueta, ***s*** los valores calculados por la red y ***w*** los pesos de la red. Considere que las entradas a la segunda capa se calculan como

$s_k = \sum_{j=0}^{n_H} x_j w_{kj} = w_k^T x_k$ donde el vector ***x*** representa la salida de la capa oculta.

a) Deducir con todo detalle la regla de adaptación de los pesos entre la capa oculta y la salida.
b) Deducir con todo detalle la regla de adaptación de los pesos entre la capa de entrada y la capa oculta.

Usar ***θ*** para notar la función de activación.

La estructura de la red ***h*** es:

$x^{(0)} \rightarrow s^{(1)} = W^1 x^{(0)} \rightarrow x^{(1)} = \theta(s^{(1)}) \rightarrow s^{(2)} = W^{(2)} x^{(1)} \rightarrow x^{(2)} = I(s^{(2)}) = h(x^{(0)})$

.

El enunciado establece, $J(W) \equiv \frac{1}{2}||t - s||^2 = \frac{1}{2}\sum_{i=1}^{c} e_i$ donde $e_i = (t_i - s_i)^2$

, $s = Wx^T$**,** $W = (W^{(1)}, W^{(2)})$.

Entonces, $\frac{\partial J}{\partial W^{(2)}} = \sum_{i=1}^{c} \frac{\partial e_i}{\partial W^{(2)}}$ y $\frac{\partial e_i}{\partial W^{(2)}} = x^{(1)}(\delta_i^{(2)})^T = x^{(1)}(\frac{\partial e_i}{\partial s^{(2)}})^T$, con

$(\frac{\partial e_i}{\partial s^{(2)}})^T = (0,\ldots,-2(t_i - s_i),\ldots,0)$ and

$\frac{\partial J}{\partial W^{(2)}} = \left[0 \ldots -2x_1^{(1)}(t_i - s_i) \ldots 0 \;||\; \ldots \ldots \ldots \ldots \ldots \;||\; 0 \ldots -2x_{n_H}^{(1)}(t_i - s_i) \ldots 0\right]$.

Sumando las matrices para cada $\boldsymbol{e_i}$ se obtiene $\frac{\partial J}{\partial W_2}$. Para $\frac{\partial J}{\partial W^{(1)}} = \sum_{i=1}^{c} \frac{\partial e_i}{\partial W^{(1)}}$.

Donde $\frac{\partial e_i}{\partial W^{(1)}} = x^{(0)}(\delta_i^{(1)})^T = x^{(0)}(\frac{\partial e_i}{\partial s^{(1)}})^T$, y

$(\frac{\partial e_i}{\partial s^{(1)}})^T = \theta'(s^{(1)}) \otimes \left[((W^{(2)})^T \delta_i^{(2)})\right]_1^{n_H}$ sustituyendo los valores $\delta^{(2)}$ y $s^{(1)}$ se obtiene $(\frac{\partial e_i}{\partial s^{(1)}})^T$. Para encontrar $\frac{\partial e_i}{\partial W^{(1)}}$ es igual que en el caso anterior, se suman las matrices.

7- Analice las restricciones de KKT del modelo SVM-Hard-Dual y diga si todo punto verificando es un vector soporte de la solución. Argumente su decisión.

La solución del problema SVM-Hard establecen que los puntos con $\alpha^* > 0$ son los vectores soporte. Pero la condición de KKT solo establece que se verifique $\alpha^*(y_n(w^T x_n + b) - 1) = 0$. Por tanto puede haber puntos no vectores soporte, es decir $\alpha^* = 0$ verificando $y_n(w^T x_n + b) = 1$. Es fácil encontrar ejemplos en $\boldsymbol{R^2}$ con puntos redundantes sobre los márgenes del pasillo de la solución.

8- Suponga un problema de predicción $\boldsymbol{(f,\mathcal{H},A,P)}$ bien definido. Suponga que la clase de funciones $\boldsymbol{\mathcal{H}}$ es suficiente para hacer que $\boldsymbol{E_{in}} \rightarrow \mathbf{0}$ para cualquier muestra de un tamaño dado $\boldsymbol{N}$.

a) Analice las fuentes de error en el problema de ajuste de una hipótesis y diga si en esta situación hay presencia de sobreajuste y, como sería la evolución del sobreajuste cuando se va reduciendo la complejidad de la clase $\boldsymbol{\mathcal{H}}$.
b) Diga en que cambiaría su análisis cuando $\boldsymbol{N}$ el tamaño de la muestra crece o decrece.

Los datos siempre están contaminados por error (pero no sabemos cual/es). Por tanto si la clase de funciones es capaz de hacer $\boldsymbol{E_{in}} = \mathbf{0}$ para toda muestra

de tamaño ***N***, dicha clase de funciones puede sufrir de severo sobreajuste al ajustar los errores estocásticos.

a) Cuando vamos disminuyendo la complejidad de la clase $\mathcal{H}$ se produce una paulatina reducción del error estocástico y un aumento del error determinístico en el ajuste, con la disminución de la varianza y un aumento de sesgo de la clase respecto de los datos. En este compromiso de errores gana la disminución por pérdida de varianza de la clase de funciones lo que hará que el sobreajuste disminuya. Este decrecimiento de sobreajuste se mantiene hasta que la variabilidad de la clase entra en compromiso con el tamaño de la muestra, en cuyo caso el error determinístico (sesgo) comenzará a ser mayor que la disminución de variabilidad de la clase de funciones y el sobreajuste comenzará a crecer de nuevo.
b) Cuando el tamaño de la muestra ***N*** crece o decrece el análisis anterior no cambia en nada, ya que no depende de ***N***. Solo depende de ***N*** la complejidad de la clase de funciones que alcanza el error mínimo de ajuste. Lo único que podemos decir es que la complejidad de la clase que alcanza el ajuste óptimo es creciente con ***N***.

9- Considere el algoritmo Perceptron y la regla de adaptación del mismo, $w_{k+1} = w_k + y_k x_k$ cuando $sign(y_k) \neq sign(w_k^T x_k)$ y $w_{k+1} = w_k$ en otro caso. Asuma el caso de datos separables. Suponga ahora que modifica dicho algoritmo de la siguiente manera: a) en lugar de ir recorriendo el conjunto de datos para aplicar la regla de adaptación, decide adaptar el valor de w con el vector x del punto peor clasificado para ese hiperplano, es decir con maor distancia absoluta al hiperplano; b) el algoritmo itera adaptando el valor de w de forma indefinida hasta que el valor de w no cambia en dos iteraciones consecutivas. Analize la situación y diga

a) ¿Converge el nuevo algoritmo a alguna solución? Justifique su contestación
b) En caso afirmativo de convergencia, ¿qué clase de solución obtiene? Justifique los argumentos

a) - Converge seguro a una solución separable porque el nuevo algoritmo, respecto del perceptron, sólo cambia el orden de reexploración de los datos y eso no afecta a ala convergencia del perceptron. Una vez el hiperplano es separador el algoritmo va maximizando la distancia de los apuntos al hiperplano y por tanto minimizando los puntos relevantes que lo definen.

b) - La solución es la misma de un SVM ya que maximiza el margen.

10- Cuando se usa “early_stopping” en el ajuste de una red neuronal es muy posible que obtengamos el mínimo de la validación en un iteración anterior a haber entrenado la red con todos los datos del conjunto de entrenamiento. Analice esta situación y diga:

a) ¿Está en conflicto esta situación con la propiedad establecida por las curvas de aprendizaje, que nos dice que cuantos más datos usemos para entrenar un modelo mejor modelo se obtiene?
b) ¿Qué pasaría si entrenamos con todos los datos?

Justifique sus decisiones con argumentos apropiados.

a) La curva de aprendizaje nos habla de la conducta del ajuste de un modelo concreto cuando el número de datos crece. No es el caso en early-stopping ya que el modelo cambia en cada iteración. Por tanto no está en conflicto porque son cosas distintas.
b) Si una vez alcanzado el mínimo de la validación seguimos entrenando hasta agotar todos los datos de entrenamiento, estaremos sobreajustando nuestro modelo a los datos. El error de la validación estará creciendo mientras el de ajuste disminuyendo.