Using Random Fourier Features with Random Forest

Albert Ribes

Director: Lluís A. Belanche Muñoz

Computer Science

Grau en Enginyeria Informàtica

Computació

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) – BarcelonaTech

Fecha de defensa

Contents

1 Ideas generales **2**

2 Introduction **3**
2.1 Problem to solve 3
2.2 Why it is important to solve this problem 3
2.3 Project proposal 3

3 Background **3**
3.1 Machine Learning 4
3.2 Classification and Regression 4
3.3 Review de los principales modelos que existen 4
3.3.1 Decision Tree 4
3.3.2 Logistic Regression 4
3.3.3 SVM 4
3.4 Las técnicas ensembling 4
3.4.1 Bagging 4
3.4.2 Boosting 4

3.5 El bootstrap . . . 4
3.6 Las funciones kernel . . . 5
3.6.1 El kernel RBF . . . 6
3.7 Las Random Fourier Features . . . 6
3.8 Nystroem . . . 7
3.9 PCA . . . 7
3.10 Cross-validation . . . 7

4 Workflow of the project **7**
4.1 La idea general un poco desarroyada . . . 7
4.1.1 Hipótesis . . . 9
4.2 Los datasets . . . 9
4.2.1 Pen Digits . . . 9
4.2.2 Covertype . . . 10
4.2.3 Satellite . . . 10
4.2.4 Vowel . . . 10
4.2.5 Fall Detection . . . 10
4.2.6 MNIST . . . 10
4.2.7 Segment . . . 10
4.2.8 Digits . . . 10
4.3 State of the art con las RFF . . . 10

5 Experimental results **10**

6 Conclussion **10**

7 Future work **10**

8 Sustainability Report **10**
8.1 Environmental . . . 11
8.2 Economic . . . 11
8.3 Social . . . 11
8.3.1 Impacto personal . . . 12
8.3.2 Impacto social . . . 12
8.3.3 Riesgos sociales . . . 12

1 Ideas generales

El guión que me propuso LLuís es:

1. Problema que ataco
2. Por qué es importante
3. Qué propongo en mi TFG
4. Estado del arte en el problema que ataco

5. Nociones generales del tema
    - Machine Learning
    - Árboles
    - Logit
    - RFF
    - Nystroem
    - Bootstrap
    - Boosting
6. El trabajo propiamente dicho (explicar lo que voy a hacer)
7. Experimentos
8. Conclusiones y Trabajo futuro
9. Referencias
10. Apéndices

2 Introduction

2.1 Problem to solve

Todavía no se consigue suficiente precisión con el Machine Learning

2.2 Why it is important to solve this problem

Con precisión más alta se podría aplicar el machine learning en otros campos

2.3 Project proposal

Incrementar el accuracy que se puede conseguir con algunos problemas mezclando la técnica del bagging (y quizá del boosting) con el tema los RFF

Actualmente el bagging solo se usa con Decision Tree porque es muy inestable. Con lo que propongo aquí, podría ser factible usarlo con otros algoritmos más estables

3 Background

Más o menos, cada uno debería ocupar entre media y una página

3.1 Machine Learning

3.2 Classification and Regression

3.3 Review de los principales modelos que existen

3.3.1 Decision Tree

3.3.2 Logistic Regression

3.3.3 SVM

3.4 Las técnicas ensembling

3.4.1 Bagging

- Inventado por Leo Breiman
- Pretende reducir el sesgo

3.4.2 Boosting

- Adaboost (adaptive boosting)
- El siguiente estimador es más probable que contenga los elementos no no se han predicho bien en el anterior
- Se trata de ir modificando los pesos que tiene cada una de las instancias
- El entrenamiento de los modelos es secuencial, a diferencia del bagging
- Enterarme de quien lo inventó, y para qué ámbitos es útil

3.5 El bootstrap

- En bagging es bueno que los estimadores estén poco relacionados entre ellos
- Idealmente, usaríamos un dataset distinto para cada uno de los estimadores, pero eso no siempre es posible
- Una alternativa es usar un resampling con repetición sobre cada uno de los estimadores para tener datasets un poco distintos entre ellos.
- Enterarme de la cantidad de elementos distintos que se espera que queden en el subconjunto, y quizá hablar de la cantidad de aleatoriedad

3.6 Las funciones kernel

Las SVM encuentran un híper-plano que separa las instancias de un problema determinado en dos subconjuntos del espacio, y en el que cada subconjunto se identifica con las clases que se quieren discriminar. Este híper-plano busca maximizar la distancia mínima entre él mismo y las instancias (los vectores) de cada una de las clases. Para hacerlo, convierte el problema en uno de optimización.

Tenemos un conjunto de datos $D = \{\boldsymbol{\chi}, \boldsymbol{y}\}$, donde $\boldsymbol{\chi} = \{\boldsymbol{x}_1, \dots \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{y} = \{-1, +1\}^n$

Se requiere encontrar $\boldsymbol{\alpha} \in \mathbb{R}^n$ que maximice:

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{1}$$

Pero este procedimiento solamente es efectivo si se da el caso que todas las instancias de $\boldsymbol{\chi}$ son linealmente separables por un híper-plano en cada una de las dos clases.

Este no siempre es el caso, y por eso se suele realizar una transformación de los datos, que los lleven de un subespacio a otro, que normalmente tiene más dimensiones que el original y que se espera que sí permita separar linealmente los datos.

Entonces, si se define una función de transformación de espacio $z(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^D$, la función a optimizar sería:

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j z(\boldsymbol{x}_i)^T z(\boldsymbol{x}_j) \tag{2}$$

Estos cálculos únicamente trabajan con el producto escalar de los vectores, nunca con ellos directamente. Es por eso que si existiera una función:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = z(\boldsymbol{x})^T z(\boldsymbol{y}) \tag{3}$$

Se podría optimizar la función

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{4}$$

sin tener jamás que calcular el producto escalar de los vectores. De hecho, la dimensionalidad del nuevo espacio vectorial podría ser infinita sin ningún problema. Lo único necesario sería que en ese nuevo espacio, que no tenemos por qué conocer, los vectores fueran linealmente separables.

Pues estas funcions κ existen, y se suelen llamar funciones kernel. Se usan especialmente con las SVM, pero se podrían usar en cualquier otro campo.

Algunas de las que existen son el kernel lineal ($\kappa(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} + c$), el polinómico ($\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + c)^d$), el gausiano o RBF ($\kappa(\boldsymbol{x}, \boldsymbol{y}) = exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2})$), etc.

3.6.1 El kernel RBF

El kernel RBF es una familia de funciones kernel. El nombre viene de *Radial Basis Function*. Esta familia de funciones tiene un parámetro σ, y son estas:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}; \sigma) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{2\sigma^2}} \tag{5}$$

A veces también lo expresan con una gamma (γ), con la equivalencia $\gamma = \frac{1}{2\sigma^2}$:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}; \gamma) = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|} \tag{6}$$

Tiene una cierta noción de similaridad entre los dos vectores: cuanto más distintos son (cuanto mayor es la norma de su diferencia) más se aproxima a 0, y si son iguales es 1.

Se sabe que el feature space de este kernel tiene dimensionalidad infinita, y es de los kernel más utilizados.

(Me gustaría enterarme si siempre siempre siempre es dimensionalidad infinita, con cualquier valor de gamma).

Un valor de σ muy pequeño (muy cercano a 0) produce más sobre-ajuste, mientras que un valor más grande lo disminuye.

3.7 Las Random Fourier Features

$$\kappa(\lambda) \approx \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle \tag{7}$$

$$\omega_i \sim \kappa(\omega) \tag{8}$$

$$\phi(x) = \frac{1}{\sqrt{D}} \left[e^{-i\omega_1^\intercal x}, \ldots, e^{-i\omega_D^\intercal x} \right]^\intercal \tag{9}$$

Es una ténica que permite aproximar el feature space de un kernel. Sea κ un kernel, tal que

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = z(\boldsymbol{x})^T z(\boldsymbol{y}) \tag{10}$$

(Creo que no permite aproximar todos los kernel, solo los que cumplen una condición)

Donde $z(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^D$. En el caso particular del kernel RBF, $z(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^\infty$

Las Random Fourier Features permiten generar una función $f(\boldsymbol{x})$ que aproxima $z(\boldsymbol{x})$ con una dimensionalidad arbitraria, de manera que $f(\boldsymbol{x})f(\boldsymbol{y}) \approx \kappa(\boldsymbol{x}, \boldsymbol{y})$

Como el subespacio de $z(\boldsymbol{x})$ es de dimensionalidad infinita para algunos kernels como el RBF, $f(\boldsymbol{x})$ coje un subconjunto aleatorio de todas esas dimensiones, según la cantidad que se haya especificado. Esto permite generar varias imágenes aleatorias de distintas aproximaciones $f(\boldsymbol{x})$ para un mismo vector $\boldsymbol{x}$, y esto mismo es lo que se explota en este trabajo para generar aleatoriedad en los datos

3.8 Nystroem

Sobretodo en el ámbito de las SVM se utiliza el concepto de *Gramm matrix* de un kernel entrenar un modelo. Sea $\boldsymbol{\chi} = \{\boldsymbol{x}_1, \dots \boldsymbol{x}_n\}$ un conjunto de datos y $\kappa(\boldsymbol{x}, \boldsymbol{y})$ un función kernel. La matriz de Gram G es de tamaño $n \times n$, y $G_{i,j} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$

El cálculo de esta matriz es muy costoso en tiempo y en espacio, y por lo tanto no es factible para la mayoría de problemas de Machine Learning, que requieren grandes cantidades de datos.

El método Nystroem consiste en aproximar esta matriz de Gram con un subconjunto aleatorio de los datos que sea adecuado sin afectar negativamente la precisión de la solución

3.9 PCA

3.10 Cross-validation

4 Workflow of the project

4.1 La idea general un poco desarroyada

Las funciones kernel son funciones que se pueden expresar de la forma:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y}) \tag{11}$$

Es decir, como producto escalar de una función de sus parámetros. Un kernel muy popular es el RBF (*Radial Basis Function*) gausiano, que es este:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}; \sigma) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{2\sigma^2}} \tag{12}$$

La función implícita ϕ ($\mathcal{L} \mapsto \mathcal{H}$) de este kernel tiene una dimensionalidad infinita, y se sabe que para cualquier conjunto de datos se puede encontrar un kernel RBF κ tal que su función implícita ϕ es capaz de separarlos mediante un híper-plano.

A pesar de que la función ϕ tiene dimensionalidad infinita ($\mathcal{H} \equiv \mathbb{R}^\infty$) es posible extraer una aproximación aleatoria de la misma con una precisión arbitraria, mediante el uso de *Random Fourier Features* [2] (RFF). Otra técnica que también se puede utilizar es el método Nystroem. Con estos métodos se puede extraer $\psi(\boldsymbol{x}) \approx \phi(\boldsymbol{x})$ y usarlos para lo que haga falta.

La extracción de estas aproximaciones se ha usado con anterioridad junto con métodos de redes neuronales, y ha mostrado muy buenos resultados. En este nosotros tratamos de usarlas con otros modelos. En particular, hemos estudiado los modelos de Decision Tree, Logit y LinearSVC, en combinación con varios tipos de ensemble.

El uso de ensembles está muy extendido junto con los Decision Tree. Esto se debe a que éste es un modelo muy inestable, y una pequeña alteración en los datos puede producir resultados muy distintos. Estas condiciones son idoneas

para hacer un comité de Decision Trees, entrenarlos con datos ligeramente distintos y elegir la solución qué más árboles hayan predicho.

Pero este procedimiento no tiene ningún sentido hacerlo con modelos que no son inestables. Si los modelos no son inestables, la mayoría de los estimadores responderán la misma solución, y no servirá para nada haber entrenado tantos modelos distintos. Es como si en un comité de expertos todos ellos opinaran igual: para eso no necesitamos todo un comité, con un solo experto nos habría bastado.

La técnica de *bagging* utiliza el *bootstrap* para generar datasets distintos para cada uno de los estimadores. Consiste en hacer un remuestreo de los datos con repetición para cada uno de los estimadores. Esta diferenciación que se produce es suficiente para los Decision Tree, pero es demasiado leve con los métodos más estables, como Logit y LinearSVC.

Pero los RFF y Nystroem abren una nueva puerta. Puesto que son aproximaciones aleatorias de un conjunto infinito, podemos sacar tantos mapeos distintos como queramos de los datos originales, y por lo tanto podemos diferenciar todavía más los datasets generados para cada uno de los estimadores.

Además de todo esto, hay una ventaja adicional: entrenar una *Support Vector Machine* (SVM) con kernel lineal es más barato que entrenar una no lineal, por ejemplo una que use RBF. Si usamos una SVM lineal, pero en vez de entrenarla con los datos originales la entrenamos con los datos $\psi(\boldsymbol{x})$, tenemos un coste similar al de entrenar una SVM lineal pero con una precisión equiparable a una RBF-SVM. Esto ya se ha hecho antes.

Existen varias formas de combinar las RFF con los métodos ensembles. Básicamente, hay dos parámetros que podemos elegir: qué tipo de ensemble usar y en qué momento usar las RFF.

Cuando se combina un ensemble con los RFF, básicamente hay dos momentos en los que se puede usar el mapeo. Un momento es nada más empezar, antes de que el ensemble haya visto los datos, y el ensemble trabaja normalmente, solo que en vez de recibir los datos originales recibe un mapeo de los mismos. Este método se abstrae completamente de lo que hace el ensemble, y lo trata como una caja negra. *Black Box*.

El otro método consiste en usar el RFF, no nada más empezar y el mismo para todos los estimadores, sino justo antes del estimador: se hace un mapeo nuevo para cada uno de los estimadores. Este método ya se mete dentro de lo que es un ensemble, y por tanto diremos que es de caja blanca (*White Box*).

Pero se sabe que se obtienen mejores resultados cuando hay bastante diversidad entre los estimadores del ensemble. Se nos presentan ahora dos formas de crear diversidad en el ensemble. Una de ellas es la forma clásica, mediante el bootstrap, que ha mostrado muy buenos resultados con el *Decision Tree*. Pero ahora podemos usar también la aleatoriedad de los RFF para generar esa diversidad. Entonces tenemos dos opciones: usar los RFF ellos solos o usarlos junto con el Bootstrap. A usarlos junto con el bootstrap le llamaremos un *Bagging*, mientras que si no usamos bootstrap le llamaemos un *Ensemble*.

Tenemos entonces varias combinaciones entre manos:

Black Bag Black Box model con Bagging. Primero se hace un mapeo de los datos y después se hace un bootstrap con ellos para cada uno de los estimadores. Si los estimadores son *Decision Tree* es los mismo que un *Random Forest*, pero no con los datos originales, sino con el mapeo.

White Bag White Box model con Bagging. Primero se hace un bootstrap de los datos, para cada uno de los modelos, y después para cada uno de ellos se hace un mapeo de los datos.

White Ensemble White Box model sin baging. Se hace un mapeo para cada uno de los estimadores, todos ellos usando todos los datos originales.

El *Black Ensemble* no tiene ningún sentido hacerlo, porque en ese caso todos los estimadores recibirían exactamente los mismos datos, y por lo tanto todos producirían exactamente los mismos resutados, a no ser que tuvieran algún tipo de aleatoriedad, como los Decision Tree. A pesar de que tengamos ese caso particular con los DT, no lo vamos a tratar.

Y luego, por supuesto, haremos pruebas con un modelo simple usando los RFF, sin usar ningún tipo de ensemble.

4.1.1 Hipótesis

De precisión:

- Usar bootstrap con RFF es demasiada aleatoriedad y producirá peores resultados que usar RFF ellos solos.
-

De tiempos:

- Podemos aproximar la precisión que tendría una RBF-SVM usando una SVM lineal con el truco de las RFF con un tiempo mucho mejor

4.2 Los datasets

He enfocado el trabajo únicamente con problemas de clasificación. He hecho pruebas con 8 datasets distintos.

Todos ellos los he normalizado a media 0 y varianza 1, y he usado dos tercios para train y un tercio para test.

4.2.1 Pen Digits

[Vease 1] Distinguir entre los 10 dígitos (0-9) de un conjunto de imágenes. El dataset se ha generado cogiendo las coordenadas x e y del trazo hecho por una persona para dibujar ese número e interpolando 8 puntos normalmente espaciados en todo el trazo del dibujo.

4.2.2 Covertype

4.2.3 Satellite

4.2.4 Vowel

4.2.5 Fall Detection

4.2.6 MNIST

4.2.7 Segment

4.2.8 Digits

4.3 State of the art con las RFF

5 Experimental results

6 Conclussion

De momento, parece que algunos problemas sí que se benefician de esto, mientas que otros no lo hacen

7 Future work

- El trabajo se ha centrado en problemas de clasificación, pero no hay ningún motivo para que no se pueda aplicar el regresión. Se ha omitodo por simplificar
- Aquella teoría de que quizá se puede regular la cantidad de aleatoriedad que añade el bootstrap, y quizá inventar un bootstrap con un parámetro para regular la cantidad de aleatoriedad
- Pensar en aquella teoría de que quizá se puede inventar un procedimiento para, dato un problema determinado con sus datos, sacar un número que sea representativo de la cantidad promedio de ruido que tiene. Puesto que quizá es útil para este proyecto conocer la cantidad de aleatoriedad que tienen los datos, para que se pueda regular

8 Sustainability Report

Consumo del equipo CO2 Materiales Peligrosos Materiales que vienen de zonas de conflicto Qué sabemos de nuestros proveedores, y si la fabricación de la maquinaria se ha hecho en una instalación sin riesgo, ni para las personas ni para la naturaleza

El impacto que tiene el proyecto directa e indirectamente en la gente Ha sido diseñado pensando en cómo se ha de reciclar y reparar, siguiendo conceptos de economía circular?

	Project development	**Exploitation**	**Risks**
Environmental	Consumption design	Ecological Footprint	Environmental risks
Economic	Project bill	Viability plan	Economic risks
Social	Personal impact	Social impact	Social risks

8.1 Environmental

El impacto medio-ambiental de mi TFG ¿Cuál es el coste medio-ambiental de los productos TIC? ¿Cuantos recursos son necesarios para fabricar un dispositivo? ¿Cuánto consumen estos dispositivos durante su vida útil? ¿Cuantos resuduos se generan para fabricarlos? ¿Qué hacemos de los dispositivos una vez ha terminado su vida útil? ¿Los tiramos y contaminamos el medio de nuestro entorno? ¿Los enviamos al tercer mundo, y contaminamos el suyo?

La huella ecológica se puede medir, por ejemplo, en Kilovatios hora y en emisiones de CO2

La huella ecológica que tendrá el proyecto durante toda su vida útil ¿Mi TFG contribuirá a reducir el consumo energético y la generación de resudios? ¿O los incrementará?

8.2 Economic

Costos de portarlo a terme i assegurar la seva pervivència

Estimar los recursos materiales y humanos necesarios para la realización del proyecto. Sería como preparar la factura que se le pasará al cliente, teniendo en cuenta que se terminará en un término bien definido

Una planificación temporal y Una explicación de si he pensado algún proceso para reducir costes

Estimar las desviaciones que he tenido respecto a las planificaciones iniciales.

Para evaluarlo durante su vida útil, he de hacer un pequeño estudio de mercado, y analizar en qué se diferencia de los ya existentes, si se mejora en algún aspecto, o no. Reflexionar sobre si mi producto tendrá mercado o no, y sobretodo, explicarlo

Además de estudiar los costes, plantearse si es posible reducirlos de alguna manera

Plantear posibles escenarios que por razones de limitaciones en el tiempo y de recursos no tendré en cuenta, pero que podrían perjudicar la viabilidad económica del proyecto

Suele salir a unos 20000 € el TFG

8.3 Social

Implicaciones sociales, tanto sobre el colectivo al que se dirige el proyecto como sobre otros colectivos

8.3.1 Impacto personal

En qué me ha afectado a mí, y a mi entorno cercano, la realización de este proyecto. En qué me ha cambiado la vida, o si ha cambiado mi visión sobre el tema. ¿Ha hecho que me de cuenta de situaciones que antes ignoraba?

8.3.2 Impacto social

Implicaciones que la realización de este proyecto puede tener sobre la sociedad. Hay que identificar al colectivo de los afectados. Los colectivos pueden ser: los propietarios, los gestores del proyecto, los trabajadores, los proveedores, los consumidores (usuarios directos), o terceros (usuarios indirectos o pasivos) Puedo consultar los estándares del GRI

8.3.3 Riesgos sociales

Explicar posibles escenarios probables, pero no significativos, que no puedo abordar por falta de tiempo o de recursos o de capacidad, y que podrían perjudicar a personas relacionadas con mi proyecto

References

[1] E. Alpaydin and Fevzi. Alimoglu. *Pen-Based Recognition of Handwritten Digits*. July 1998. URL: `https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits`.

[2] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: (2007).