UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE MATEMÁTICA

SEMINARIO DE TESIS DE MATEMÁTICA PURA Y APLICADA II

REGRESIÓN

por

PROCESOS GAUSSIANOS.

Un proyecto presentado por Anthony Enrique Huertas Quispe

Supervisado por:
Edgard Kenny Venegas Palacios

Lima, Perú
2016

Este trabajo esta dedicado
en especial a mi madre quien
confía en mis éxitos profesionales.

Índice general

Introduction III

1. Machine Learning 1
- 1.1. Aprendizaje Supervisado . . . 2
  - 1.1.1. Regresión . . . 3
    - 1.1.1.1. Algoritmo principal de Regresión . . . 3
  - 1.1.2. Clasificación . . . 3
    - 1.1.2.1. Algoritmos principales de Clasificación . . . 4
- 1.2. Aprendizaje no Supervisado . . . 6
  - 1.2.1. Agrupamiento (Clustering) . . . 6
    - 1.2.1.1. Algoritmos principales de Clustering . . . 7
  - 1.2.2. Estimación de densidad . . . 7
    - 1.2.2.1. Algoritmos principales de Estimación de densidad . . . 7
  - 1.2.3. Reducción de Dimensionalidad . . . 8
    - 1.2.3.1. Algoritmos principales de Reducción de Dimensionalidad . . . 8
- 1.3. Aprendizaje por Refuerzo . . . 9

2. Nociones Básicas 10
- 2.1. Álgebra lineal . . . 10
  - 2.1.1. Matrices y Vectores . . . 10
  - 2.1.2. Eigenvectores y Eigenvalores . . . 13
  - 2.1.3. Algunas funciones . . . 14
- 2.2. Estadística y Probabilidades . . . 14
  - 2.2.1. Elementos de Probabilidad . . . 14
  - 2.2.2. Variables aleatorias . . . 15
  - 2.2.3. Independencia y Condicionalidad . . . 15
  - 2.2.4. Probabilidad Bayesiana . . . 16
  - 2.2.5. Procesos estocásticos . . . 16
  - 2.2.6. Distribuciones . . . 17
  - 2.2.7. Tendencias centrales y dispersión . . . 20
    - 2.2.7.1. Estadística descriptiva . . . 22

2.2.8. Distribución Gaussiana (Multivariable) . . . 24
2.2.8.1. Intervalos de confianza . . . 25
2.2.9. Estadística Inferencial . . . 26
2.2.9.1. Principio de Verosimilitud . . . 26
2.2.9.2. Estimador de Máxima Verosimilitud (E.V.M.) . . . 26
2.2.9.3. Estimador de Máximo a Posteriori (M.A.P.) . . . 27
2.3. Optimización . . . 27
2.3.1. Convexidad . . . 27
2.3.2. Descenso del Gradiente . . . 28

3. Regresión Lineal **30**
3.1. Regresión Lineal Simple . . . 30
3.1.1. Función Costo (Minimización) . . . 32
3.2. Regresión lineal múltiple . . . 33
3.2.1. Función Costo (Minimización) . . . 34
3.3. Análisis Bayesiano . . . 34

4. Procesos Gaussianos **36**
4.1. Motivación . . . 36
4.2. Análisis Constructivos . . . 39
4.2.1. Primer Análisis (Distribución predictiva) . . . 39
4.2.2. Segundo Análisis (Extensión de la Dimensionalidad) . . . 41
4.2.3. Tercer Análisis (Función kernel) . . . 44
4.3. Regresión por GP's . . . 45
4.3.1. Distribución Predictiva Conjunta . . . 47
4.3.2. Entrenamiento del Modelo (Regresión) . . . 49
4.4. Hiperparámetros . . . 50
4.4.1. Maximá verosimilitud . . . 51
4.5. Algoritmo GPR . . . 52

5. Conclusiones y recomendaciones **53**

Introducción

Las técnicas de *Regresión* representan parte de estudio en el campo de predicción tomando como base resultados numéricos de una serie de causas también numéricas denominadas experiencias, y como objetivo enfocar nuevos resultados a causa de ellas; por ejemplo, se puede ajustar estas técnicas para predecir precios del mercado tomando como base los precios anteriores y las características que generaron tales resultados.

El campo de la predicción forma parte de un amplio campo denominado *Machine Learning*, el cual esta siendo desarrollado desde el origen de la tecnología de información, y va creciendo con la implementación de mejoras o diseño de nuevas técnicas ya sea generalizando objetivos o estableciéndose particularidades según el aréa en el cual se desee trabajar.

Este proyecto va enfocado directamente a la mejora de la técnica de regresión mediante *Procesos Gaussianos*, estableciéndonos análisis estadísticos en un espacio de funciones, y otorgándonos un alto porcentaje de certeza sobre un conjunto pequeño de posibles resultados facilitando la toma de decisiones en cuanto a un valor de predicción.

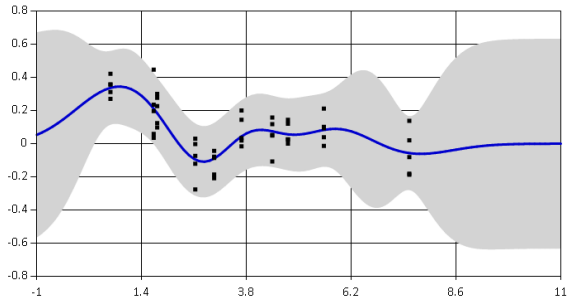


Figura 1: -Puntos: Valores reales observadores. -línea azul: Predicción de valores. -Zona gris: zona de confianza.

Con el objetivo de enmarcar nuestras ideas hacia un estudio que toma como base ciertos resultados matemáticos; en este trabajo se implementan los conpectos básicos respecto a ciertas áreas.

Capítulo 1

Machine Learning

En esta sección, el objetivo será comprender el significado de **Machine Learning**, o aprendizaje automatizado, tanto desde una visión matemática como computacional. En primera instancia, Christopher Bishop, subgerente en el laboratorio de investigación de Microsoft en Cambridge, enunció lo siguiente:
"Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field...", en español, *"Los patrones de reconocimiento tienen sus orígenes en la ingeniería, mientras que el aprendizaje automatizado proviene de la ciencia de la computación. Sin embargo, estas actividades pueden ser vistas como dos facetas de un mismo campo..."*
lo cual nos da a entender que, desde su perspectiva, los métodos que se desarrollaron en la ciencia de la computación aprovechan métodos semejantes a los que utiliza la ingeniería y viceversa; sin embargo, si la Informática o tecnología de la información satisface nuestro requerimiento eso es llamado aprendizaje automatizado.

Por otra parte, Arthur Samuel, pionero americano en el campo de aprendizaje automatizado, lo define como:
"Field of study that gives computers the ability to learn without being explicitly programmed", en español, *"Campo de estudio que da a las computadoras la habilidad de aprender sin ser programadas explícitamente."*
Arthur inició estos estudios cuando decidió enseñar a una máquina llamada *Defense Calculator* a jugar el famoso "*Juego de Damas*"; el análisis implementado fue de tal modo que el sistema no necesite calcular el número de jugadas posibles y optar por la mejor, pues resultaría una complejidad, sino que logre hacer un conteo de piezas y mediante instrucciones básicas ejecute la que genere como resultado más piezas de ventaja sobre el adversario; esto revolucionó el razonamiento científico pues hoy en día un aprendizaje automatizado facilita la complejidad de un resultado.

La idea que se opta hoy en día es la que fue enunciada por Tom Mitchell sobre el

problema de aprendizaje bien planteado:
"*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*", en español, "*Se dice que un programa de computadora aprende de una experiencia E con respecto a alguna tarea T y una medida de rendimiento P, si el rendimiento en la tarea T, con uso de la medida P, mejora con la experiencia E.*"
Un ejemplo sencillo, es cuando un sistema aprende como filtrar correos Spam, en este caso la experiencia E sería la observación sobre los correos que uno como usuario decidió marcarlas como tal, la tarea T del sistema sería la de clasificar correos como Spam o no en base a la experiencia E, y la medida de rendimiento P vendría a ser el número de correos que correctamente han sido clasificados como Spam.
En este capítulo, daremos a conocer algunas técnicas de aprendizaje automatizado, en donde los dos primeros son los más sobresalientes para este proyecto.

1.1. Aprendizaje Supervisado

El aprendizaje supervisado (*Supervised Learning*) es implementado en un sistema fundamentando su utilidad en el campo de la predicción, basando sus experiencias en características que generan efectos reales, valores de salida (numéricos o etiquetas), mediante un *algoritmo de Machine Learning*, y aprende de ellas para la elaboración de un modelo; consiguientemente con un *algoritmo de predicción* basado en nuestro modelo generamos los valores de salida predichos para nuevos datos (datos de prueba); en otras palabras, se enseña a un programa cómo realizar una tarea y que haga uso de aquel conocimiento adquirido para problemas futuros. Las aplicaciones son multivariadas, pero clasificadas por cómo se enfoquen dichos efectos.

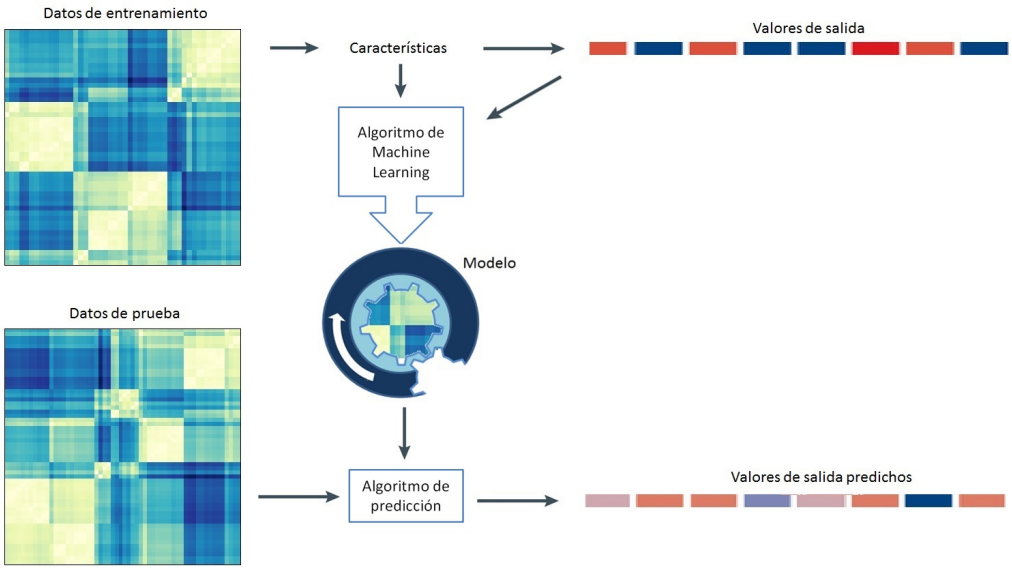


Figura 1.1: Aprendizaje Supervisado (Modelo Gráfico).

1.1.1. Regresión

Esta técnica es efectuada en un sistema para predecir uno o múltiples valores numéricos de salida de datos en base a sus características, y a la experiencia adquirida sobre un conjunto de datos con sus valores de salida reales. Un ejemplo de aprendizaje automatizado por regresión, es el de predecir el progreso de la enfermedad de diabetes en base a variables psicológicas (edad, sexo, peso, presión sanguínea).

1.1.1.1. Algoritmo principal de Regresión

- **Regresión Lineal:** La técnica de regresión lineal, es muy útil y sencilla de elaborar pues establece un modelo lineal de la forma $y = \mathbf{x}^T\theta$ que nos permite relacionar linealmente los valores de salida numéricos dependientes de las características de los datos. A continuación, se visualiza el diseño de un modelo de regresión lineal para 442 pacientes sobre el ejemplo de diábetes planteado anteriormente; la base de datos se encuentra en la librería scikit-learn.datasets.

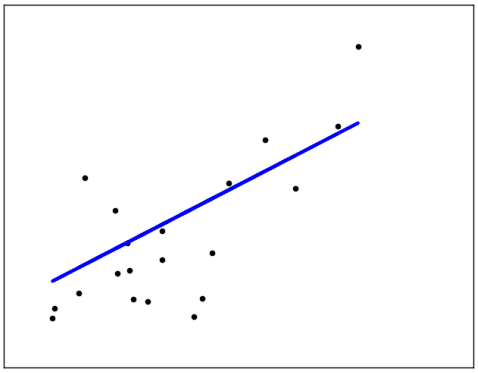


Figura 1.2: Regresión Lineal.

1.1.2. Clasificación

Consiste en enseñar a un sistema que, bajo experiencias en datos que ya se encuentran clasificados de acuerdo a ciertas características, continúe un proceso de clasificación con nuevos datos. Un ejemplo de un sistema que realice esta tarea, es el de contar con la longitud y ancho de sépalos de flores iris que se encuentran clasificadas como Setosa, Versicolor o Virgínica, y en base a ello continuar la clasificación para nuevos datos con tales características.

1.1.2.1. Algoritmos principales de Clasificación

- **Regresión Logística**[1]**:** A pesar del término "regresión", este algoritmo de clasificación no corresponde a la técnica de Regresión, propiamente dicho. El análisis mediante regresión logística es la de establecer una relación entre las características de ciertos datos y valores discretos, permitiéndo modelar un sistema que prediga consiguientemente este tipo de relaciones para nuevos datos. A continuación se visualiza la clasificación de datos con los valores discretos $-\mathbf{1}$ y $\mathbf{1}$.

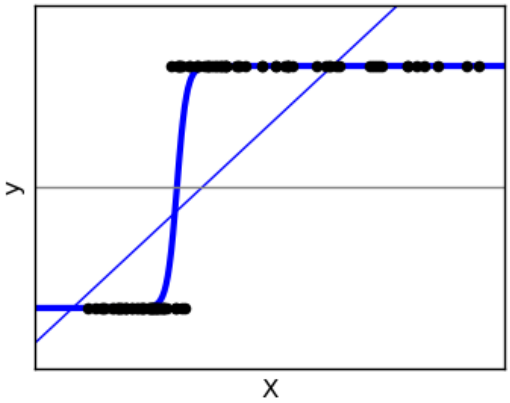


Figura 1.3: Regresión Logística.

- **Árboles de Decisión:** Esta técnica clasifica variables tanto continuas independientes como cualitativas[2] agrupándolas en un diagrama árbol[3], estableciendo un camino predictivo para un dato de entrada basándose en sus características.

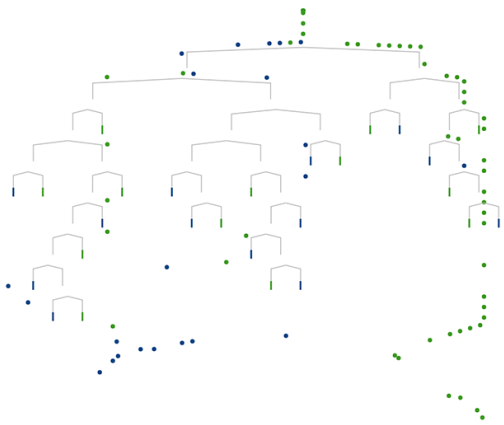


Figura 1.4: Arbol de decisión, modelado por los datos visualizados como puntos.

[1] Algunos autores establecen este algoritmo como parte de la técnica de Regresión.
[2] Variables que pueden ser medidas o presentar un orden.
[3] Diagrama estructurado por categorías.

- **Support Vector Machine (SVM)**[4]**:** Esta técnica distribuye datos $k-$ *dimensionales* en un espacio $\mathbb{R}^k$, donde cada eje de coordenadas se denomina vector soporte, para luego establecerse una geometría de separación aproximada sobre los datos. A continuación, se visualiza zonas de clasificación establecidas sobre dos características.

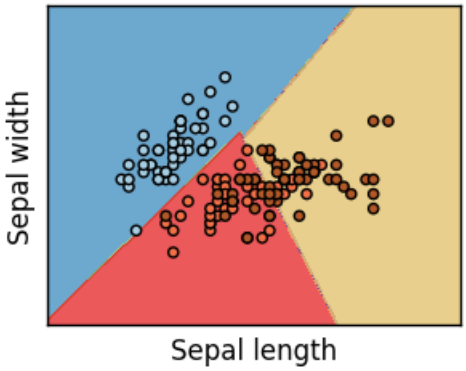


Figura 1.5: SVM.

- **Naive Bayes:** Esta técnica hace uso del Teorema de Bayes, lo cual nos establece lo siguiente:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

donde P es una función de probabilidad; la idea de este resultado es predecir "y" en base al conjunto de datos "X", lográndose esto cuando $P(y|X)$ sea el máximo valor en comparación con cualquier otro valor $P(y'|X)$.

- **K-vecinos más cercanos (KNN)**[5]**:** Esta técnica clasifica a un dato en base a las clasificación de los K datos más cercanos[6] que lo rodean.

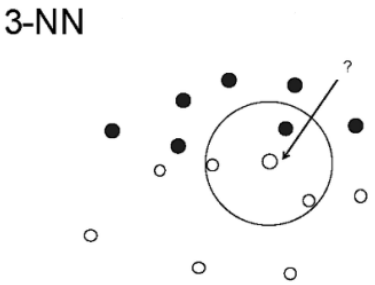


Figura 1.6: KNN.

[4]A pesar de ser una técnica de clasificación, puede implementarse en técnicas de regresión lineal.
[5]Está técnica también tiene implementaciones en técnicas de Regresión
[6]Esto refiere a la implementación de una métrica que actue sobre los valores característicos de los datos

1.2. Aprendizaje no Supervisado

El aprendizaje no supervisado *(Unsupervising Learning)* es implementado en un sistema cuando se necesita estructurar datos que no poseen valores de salida o etiquetas; en otras palabras, representa una serie de técnicas que hace que un sistema aprenda a resolver tareas en base a experiencias que son adquiridas por sí mismo.

Este tipo de aprendizaje es fundamental en la estructuración de los datos cuando no se tiene una información inicial sobre estos, a diferencia del aprendizaje supervisado, que basa sus experiencias a causa de efectos reales. Por tanto la elaboración de un *algoritmo de Machine Learning*, en este caso, estudia las características de datos (datos de entrenamiento), y las estructura en base a ciertas similitudes que se puedan encontrar en la distribución, por lo que consiguientemente con un *algoritmo de estructuración* basado en nuestro modelo, se estructuran nuevos datos (datos de prueba).

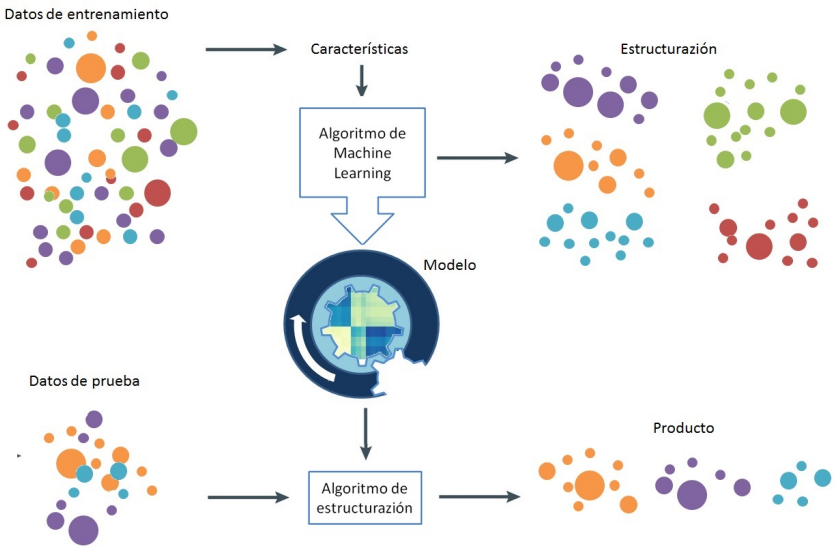


Figura 1.7: Aprendizaje no Supervisado (Modelo Gráfico).

1.2.1. Agrupamiento (Clustering)

A esta técnica del aprendizaje no supervisado se le denomina *Clustering* y agrupa datos que posean similitud en base a la distribución de sus características logrando diferenciar los datos unos de otros. De forma similar a los algoritmos de *Clasificación*, Clustering logra en ciertos aspectos clasificar datos; sin embargo, el proceso es sin tener un conocimiento a priori de algún valor de salida ni etiquetas sobre los datos.

1.2.1.1. Algoritmos principales de Clustering

- **K-means:** Esta técnica asume K agrupamientos (clusters), regidos bajo una característica central, sobre los datos en donde cada una de ellas mantengan cierta homogeneidad con respecto a evaluar datos pertenecientes a distintos grupos, manteniéndose heterogeneidad entre grupos. La idea es obtener el K óptimo que satisfaga lo explicado anteriormente.

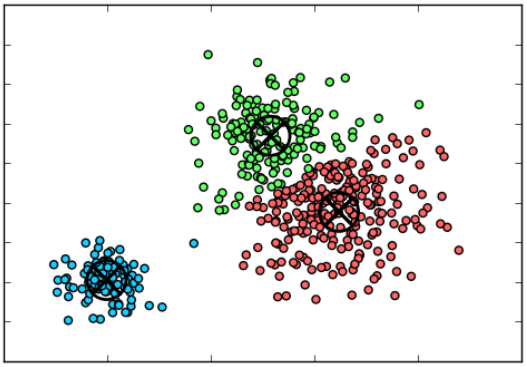


Figura 1.8: K-means.

1.2.2. Estimación de densidad

Este método basa su construcción en un estimador que modele datos observados, formando distribuciones aproximadas.

1.2.2.1. Algoritmos principales de Estimación de densidad

- **Estimación de Densidad de Kernel (KDE):** Esta técnica estima la función de probabilidad de una variable, en este caso de dimensión (característica) de los datos mediante estadística inferencial en base a una cantidad de datos finito.

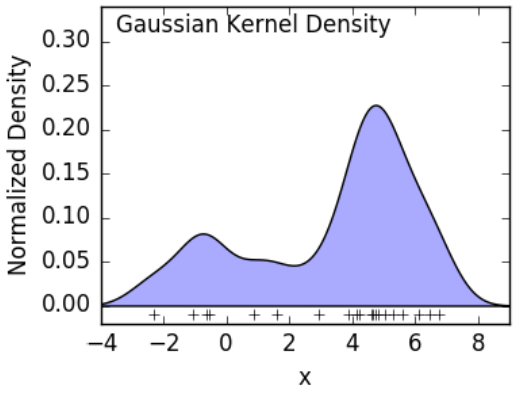


Figura 1.9: Estimación de densidad.

1.2.3. Reducción de Dimensionalidad

El uso correspondiente de esta técnica tiene una alta importancia al trabajar en dimensiones muy elevadas pues dado datos con un número alto de carácterísticas, la reducción de dimensionalidad procesa el aprendizaje basando su estudio en una dimensión reducida. A continuación, se visualizará un conjunto de datos, en donde, en una dirección, puede ser representando como una nube de puntos tridimiensional mientras, que en otra, se aproxima a un plano bidimensional; he aquí una muy buena aplicación de la reducción de dimensionalidad.

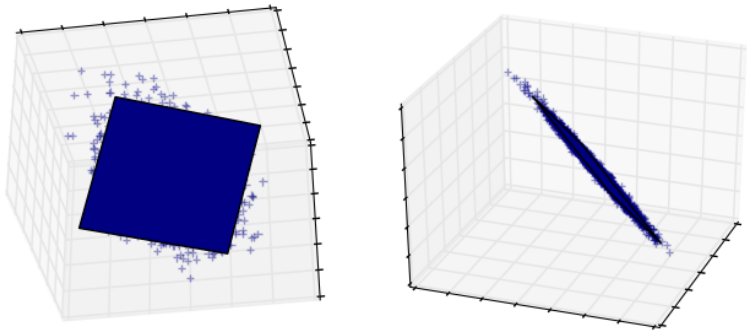


Figura 1.10: Reducción de dimensionalidad.

1.2.3.1. Algoritmos principales de Reducción de Dimensionalidad

- **Descomposición de Valores Singulares (SVD)[7]:** Está técnica en particular es un resultado propio del álgebra lineal; sin embargo, es importante representarlo como una técnica de reducción de dimensionalidad pues por su construcción reduce cálculos de dimensiones altas a dimensiones de menor valor generando un menor costo computacional. Esta técnica descompone una matriz $A \in \mathbb{R}^{m \times n}$ como $A = U\Sigma V^T$ donde U, V son autovectores de las matrices AA^T y A^TA respectivamente, y Σ matriz diagonal que mantiene los valores singulares de A.

- **Análisis de Componentes Principales (PCA):** El procedimiento llevado por esta técnica es la de establecer un subespacio de menor dimensión que los datos, tal que la pérdidad de información sobre los datos establecida en dicho subespacio, sea insignificante.

[7]Existen múltiples técnicas al igual que SVD que reducen de forma similar el costo computacional como lo es la descomposición QR, Cholesky, LU, etc.

1.3. Aprendizaje por Refuerzo

El aprendizaje por refuerzo *(Reinforcement Learning)* es uno de los más importantes en su uso sobre el campo de la inteligencia artificial. Esta técnica de aprendizaje es mucho más compleja pues implementa una relación **recompensa - castigo** en su proceso, es decir, que mientras el sistema desarrolle una tarea y se vea bien realizada se le recompensará asignándole un valor positivo, en caso contrario se le castigará, asignándole un valor negativo; por lo que nuestro sistema mejora rendimientos, analiza situaciones y realiza nuevas tareas quizás desarrolladas por sí mismo.

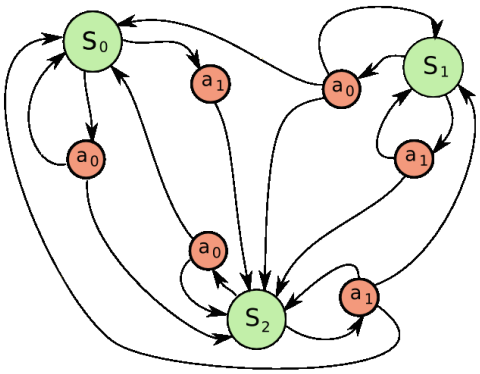


Figura 1.11: Aprendizaje por Refuerzo (Modelo Gráfico).

Capítulo 2

Nociones Básicas

Con el fin de contar con las herramientas básicas para el análisis correspondiente de una técnica de Regresión por Procesos Gaussiano, se definirán algunos conceptos del álgebra lineal, estadística descriptiva e inferencial y optimización que influyen adecuadamente.

2.1. Álgebra lineal

Evitando la complejidad de los resultados, el espacio en el que se enfocará es en R^n, por tanto los conceptos dados se limitan a este espacio.

2.1.1. Matrices y Vectores

Definición 2.1. (Matriz - vector - dimensión). *Denominaremos* **matriz** *a un conjunto de números reales agrupados en n filas y m columnas, denotado como* $A \in \mathbb{R}^{n\times m}$*, siendo* $n \times m$ *su* **dimensión***, y presenta la siguiente estructura*

$$A = \begin{bmatrix} a_{11} & a_{12} \dots & a_{1m} & \\ a_{21} & \ddots & \ddots & a_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \dots & \dots & a_{nm} \end{bmatrix} = [a_{ij}]_{n\times m}$$

Se denominará **vector** *cuando* $m = 1$ *y se denota como* $v \in \mathbb{R}^n$*. Su representación es*

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Operaciones con matrices:

- Suma: $A + B = [a_{ij} + b_{ij}]_{n\times m}$ donde $A = [a_i j]_{n\times m}$ y $B = [b_i j]_{n\times m}$.
- Producto: $AB = \left[\sum_{k=1}^{p} a_{ik}b_{kj}\right]_{n\times m}$ donde $A = [a_i j]_{n\times p}$ y $B = [b_i j]_{p\times m}$.
  Este producto solo está definido si el número de columnas de A es el número de filas de B.
- Producto: $cA = [ca_{ij}]_{n\times m}$ donde $A = [a_i j]_{n\times p}$ y $c \in \mathbb{R}$.

Propiedades del producto de matrices:

- Asociatividad: $(AB)C = A(BC)$
- Distributividad: $A(B + C) = AB + AC$
- No Conmutatividad: Si $A = [a_{ij}]_{n\times p}$ y $B = [b_{ij}]_{p\times m}$, luego $AB = [c_{ij}]_{n\times m}$; sin embargo, BA no está definido si $n \neq m$. En caso $m = n$, entonces $AB = [c_{ij}]_{n\times n}$ y $BA = [c_{ij}]_{p\times p}$ pero no necesariamente $n = p$, es decir $AB \neq BA$ en general.

Tipos matrices:

- **Matriz cuadrada**: Sea $A \in \mathbb{R}^{m\times n}$, diremos que A es una matriz cuadrada si $m = n$.
- **Matriz diagonal** Sea $A \in \mathbb{R}^{n\times n}$, diremos que $A = [a_{ij}]_{n\times n}$ es una matriz diagonal si $a_{ij} = 0$ para todo $i \neq j$.
- **Matriz identidad**: La matriz identidad $\mathbb{I} \in \mathbb{R}^{n\times n}$, donde $\mathbb{I}_{n\times n} = [a_{ij}]_{n\times n}$, se define como una matriz diagonal tal que $a_{ij} = 1$ para todo $i = j$.
- **Matriz transpuesta**: Sea $A \in \mathbb{R}^{n\times n}$, donde $A = [a_{ij}]_{m\times n}$, definimos su transpuesta como $A^T = [a_{ji}]_{n\times m}$.
- **Matriz simétrica**: Sea $A \in \mathbb{R}^{m\times n}$, diremos que A es una matriz simétrica si $A = A^T$.
- **Matriz inversa**: Sea $A \in \mathbb{R}^{n\times n}$, diremos que A tiene inversa si existe una matriz $B \in \mathbb{R}^{n\times n}$ tal que $AB = I$. L matriz inversa se denota como A^{-1}.
  (Se cumple la conmutatividad $AA^{-1} = A^{-1}A = I$).
- **Matriz ortogonal.**: Sea $A \in \mathbb{R}^{n\times n}(n \geq 2)$, es matriz ortogonal si $A^{-1} = A^T$.

Definición 2.2. (Norma). *Se denota la norma de un vector $v \in R^n$ por $\|v\|$ y se define como*

$$\|v\| = \sqrt{v \cdot v} = \sqrt{\sum_{i=1}^{n} v_i^2}, \quad \textit{donde } v = [v_i]_{n\times 1},$$

representando su longitud.

Definición 2.3. (Distancia entre vectores) *Se define la distancia entre* $u, v \in \mathbb{R}^n$ *como*

$$d(u, v) = \|u - v\|.$$

Lema 2.1. (Inversión Matricial) *Sean* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$ *invertibles, y además* $U, V \in \mathbb{R}^{n\times m}$, *entonces*

$$(A + UBV^T)^{-1} \quad = \quad A^{-1} - A^{-1}U(B^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1}$$

Demostración. Representemos el siguiente sistema de ecuaciones, para resultados matriciales

$$\begin{bmatrix} A_{n\times n} & U_{n\times m} \\ V^T_{m\times n} & -B^{-1}_{m\times m} \end{bmatrix} \begin{bmatrix} X_{n\times n} \\ Y_{m\times n} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{n\times n} \\ \mathbf{0}_{m\times n} \end{bmatrix} \quad \Rightarrow \quad \begin{array}{rcl} AX + UY & = & \mathbb{I} \\ V^TX - B^{-1}Y & = & \mathbf{0} \end{array}$$

Luego, $V^TX - B^{-1}Y = \mathbf{0}$ implica que $BV^TX = Y$ y por tanto de la expresión $AX + UY = \mathbb{I}$ tenemos que

$$AX + UY = \mathbb{I} \begin{cases} AX = \mathbb{I} - UY & \Rightarrow \quad X = A^{-1}(\mathbb{I} - UY) \\ AX + U(BV^TX) = \mathbb{I} & \Rightarrow \quad X = (A + UBV^T)^{-1} \end{cases} \tag{2.1}$$

Ahora

$$\begin{aligned} V^TX - B^{-1}Y = \mathbf{0} \quad &\Rightarrow \quad V^T(A^{-1}(\mathbb{I} - UY)) - B^{-1}Y = \mathbf{0} \\ &\Rightarrow \quad V^TA^{-1} = (B^{-1} + V^TA^{-1}U)Y \\ &\Rightarrow \quad Y = (B^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1} \end{aligned} \tag{2.2}$$

Comparando (2.2) en (2.1) con $X = (A + UBV^T)^{-1}$, tenemos que

$$X = (A + UBV^T)^{-1} \tag{2.3}$$

$$\begin{aligned} X &= A^{-1}(\mathbb{I} - UY) = A^{-1}(\mathbb{I} - U((B^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1})) \\ &= A^{-1} - A^{-1}U((B^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1}) \end{aligned} \tag{2.4}$$

por tanto, de (2.3) y (2.4), se concluye la prueba.

$$(A + UBV^T)^{-1} \quad = \quad A^{-1} - A^{-1}U((B^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1})$$

□

Definición 2.4. (Ortogonalidad). *Sean* $u, v \in \mathbb{R}^n$, *se dice que* u *y* v *son ortogonales si* $u \cdot v = 0$. *Además, un conjunto* $\{u_1, u_2, \ldots, u_m\} \subset \mathbb{R}^n$ $(m \leq n)$ *es un conjunto ortogonal si* $u_i \cdot u_j = 0$ *para todo* $i \neq j$.

Definición 2.5. (Vector Normal). *Un vector $v \in \mathbb{R}^n$ es normal, o unitario, si $\|v\| = 1$.*

Definición 2.6. (Ortonormalidad). *Un conjunto $\{u_1, u_2, \ldots, u_m\} \subset \mathbb{R}^n$ $(m \leq n)$ es ortonormal si es un conjunto ortogonal y además $\|u_i\| = 1$ para todo $i = 1, \ldots, n$.*

Corolario 2.1. *Sea $A \in \mathbb{R}^{n\times n}$, si sus vectores columna $u_1, \ldots, u_n$ forman un conjunto ortonormal, entonces A es ortogonal.*

Demostración. Veamos que A es ortogonal, es decir que $A^{-1} = A^T$

$$A^T A = \begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \ldots & u_n \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{bmatrix} = \mathbb{I}_{n\times n}$$

lo último es a causa de que $u_i^T u_i = 1$ y $u_i^T u_j = 0$ si $i \neq j$. □

Definición 2.7. (Linealidad independiente) *Un conjunto $\{u_1, u_2, \ldots, u_m\} \subset \mathbb{R}^n$ $(m \leq n)$ es linealmente independiente si la siguiente igualdad*

$$\sum_{i=1}^{m} c_i u_i = 0 \quad \text{con } c_i \in \mathbb{R} \text{ para todo } i = 1, \ldots, m$$

solo se satisface cuando $c_i = 0$ para todo $i = 1, \ldots, m$.

Definición 2.8. *(***Rango.)** *Sea $A \in \mathbb{R}^{n\times m}$, se define el rango de A como el número de columnas de A linealmente independientes, y denotado por $Rang(A)$.*

2.1.2. Eigenvectores y Eigenvalores

Definición 2.9. (Eigenvector-Eigenvalor)*: Sea $A \in \mathbb{R}^{n\times n}$ y $v \in \mathbb{R}^n$ distinto de cero, se dice que v es eigenvector de A si*

$$Av = \lambda v, \quad \text{para algún } \lambda \in \mathbb{R}.$$

Se dice que λ es el eigenvalor de A, correspondiente a v.

Definición 2.10. (Valores Singulares). *Sea $A \in \mathbb{R}^{n\times m}$, se definen los valores singulares de A como las raíces cuadradas de los eigenvalores de $A^T A$.*

Definición 2.11. (Diagonalización) *Sea $A \in \mathbb{R}^{n\times n}$, se dice que A es diagonalizable si y solo si A puede ser descompuesta como*

$$A = UDU^{-1} \quad \text{donde } U = \begin{bmatrix} u_1 & \ldots & u_n \end{bmatrix} \text{ y } D = diag(\lambda_1, \ldots, \lambda_n).$$

donde $u_1, \ldots, u_n$ son eigenvectores de A, asociados a los eigenvectores $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.
Se dice que A es ortogonalmente diagonalizable si $A = UDU^T$.

2.1.3. Algunas funciones

Definición 2.12. *(Delta de Kronecker) Sean $x_i, x_j \in \mathbb{R}^n$, se define el delta de Kronecker, como*

$$\delta_{ij} = \delta_{ij}(x_i, x_j) = \begin{cases} 1 & si\ i = j. \\ 0 & si\ i \neq j. \end{cases}$$

2.2. Estadística y Probabilidades

La teoría de estadística y probabilidades que nos concierne en esta sección, nos describe las herramientas que corresponden para los análisis en los siguiente capítulos.

2.2.1. Elementos de Probabilidad

Definiremos los elementos que definen una probabilidad

- Espacio Muestral (Ω): El conjunto de salidas de un experimento aleatorio; donde cada elemento se denota por ω, denominado suceso.
- Conjunto de Eventos ($\mathcal{F}$): Subconjunto de Ω cuyos elementos, denominados eventos, representan una colección de resultados posibles de un experimento aleatorio.
- Medida de probabilidad: Una función $P : \mathcal{F} \to \mathbb{R}$ se dice que es probabilidad si satisface los siguientes axiomas, denominados **axiomas de Kolmogórov**.
  - $P(A) \geq 0$, para todo $A \in \mathcal{F}$.
  - $P(\Omega) = 1$.
  - Si $A_1, A_2, \ldots$ son eventos disjuntos, es decir, $A_i \cap A_j = \emptyset$, entonces

$$P(\cup_i A_i) = \sum_i A_i$$

Propiedades:

- $P(\emptyset) = 0$.
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Definición 2.13. (Espacio muestral finito equiprobable). *Cuando Ω es un conjunto finito de sucesos $\omega_1, \ldots, \omega_n$ con la misma probabilidad, esto es $P(\omega_i) = \frac{1}{n}$.*

Definición 2.14. (Espacio muestral finito no equiprobable). *Cuando Ω es un conjunto finito de sucesos $\omega_1, \ldots, \omega_n$ no todos con la misma probabilidad.*

Definición 2.15. (Espacio muestral infinito) *Cuando Ω es un conjunto infinito de sucesos.*

2.2.2. Variables aleatorias

Los resultados de un experimento pueden tener naturaleza tanto numérica como no numérica. El concepto de variable aleatoria es el de establecer un conjunto numérico de salidas para todo tipo de experimentos.

Definición 2.16. (Variable aleatoria). *Una variable aleatoria X con valores en el conjunto E es una función $X : \Omega \to E$.*

Un ejemplo sencillo es $\Omega = \{C, S\}$ representando los resultados del experimento al lanzar una moneda, la cual sería de naturaleza no numérica. La variable aleatoria X puede ser definida como $X(C) = 0$ y $X(S) = 1$, y por tanto el conjunto $E = \{0, 1\}$.
Además se define una **variable aleatoria discreta** si E es un conjunto finito; y una **variable aleatoria continua** si E es un conjunto infinito. De ahora en adelante haremos mención a una variable aleatoria como **R.V.**

2.2.3. Independencia y Condicionalidad

Definición 2.17. (Probabilidad Condicional). *Dado dos eventos A y B, la probabilidad de que ocurra A, luego de que B haya ocurrido, se define como*

$$P(A|B) \quad = \quad \frac{P(A \cap B)}{P(B)}, \text{ dado } P(B) > 0.$$

Definición 2.18. (Independencia). *Dado dos eventos A y B, se dice que son independientes si*

$$P(A|B) \quad = \quad P(A)$$

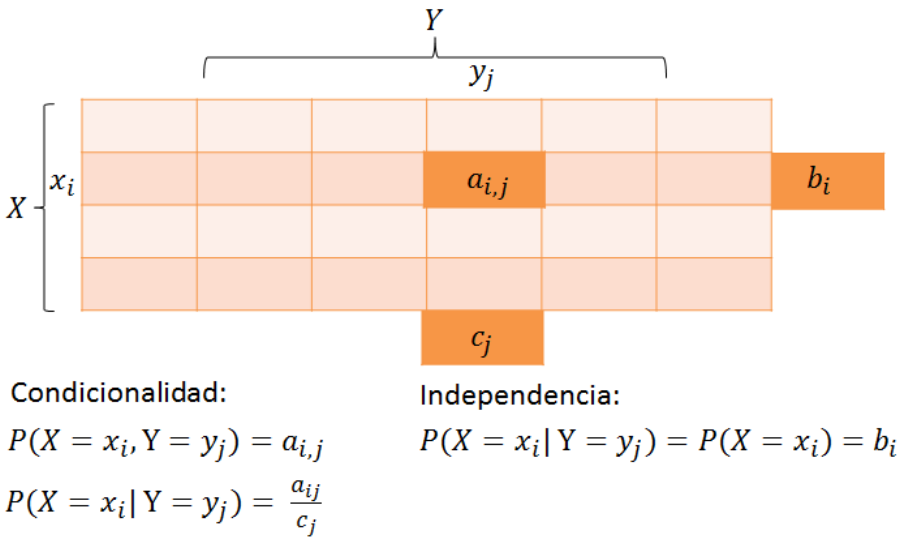


Figura 2.1: Condicionalidad, Independencia. (Variables discretas)

Propiedades:

- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$. (Condicionalidad)
- $P(A \cap B) = P(A)P(B)$. (Independencia)

2.2.4. Probabilidad Bayesiana

Teorema 2.1. (Teorema de la Probabilidad Total) *Sean $A_1, \ldots, A_n$ sucesos de Ω con $P(A_i) > 0$ para todo $i = 1, \ldots, n$ tales que*

- $A_i \cap A_j = \emptyset (i \neq j)$.
- $\Omega = \bigcup_{i=1}^{n} A_i$.

entonces,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Teorema 2.2. (Teorema Bayesiano). *Si $P(B) > 0$, con las condiciones del Teorema anterior, entonces*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

2.2.5. Procesos estocásticos

Definición 2.19. (Proceso Estocástico). *Un proceso estocástico con espacio de estado E es una colección de variables aleatorias X_t, esto es $\{X_t : t \in T\}$, definidas en un mismo espacio de probabilidad tales que $X_t : \Omega \to E$.*

El conjunto T es llamado *conjunto de parámetros*. En caso T es un conjunto contable, el proceso se denomina **proceso con parámetro discreto**; si T es un conjunto no contable, el proceso se denomina **proceso con parámetro continuo**.
Daremos un ejemplo sencillo, suponemos Ω el conjunto de posible salidas al medir los tiempos de llegadas de clientes en una tienda, es decir que un suceso $\omega \in \Omega$ es de la forma $\omega = (\omega_1, \ldots)$ donde ω_1 sería el tiempo de llegada del primer cliente y así sucesivamente. Asignando una variable aleatoria X_t como el número de llegadas en un intervalo de tiempo $(0, t]$, esto es

$$X_t(\omega) = k \text{ si y solo si } \sum_{i=1}^{k} \omega_k \leq t < \sum_{i=1}^{k} \omega_{k+1}.$$

Por tanto $X_t : \Omega \to E = \{0, 1, \ldots\}$. Luego dado que $t \in (0, t] \in \mathbb{R}_+$, entonces el proceso estocástico $\{X_t : t \in \mathbb{R}_+\}$ es un proceso con parámetro continuo, que representa el proceso de llegadas de clientes en una tienda para el respectivo ejemplo.

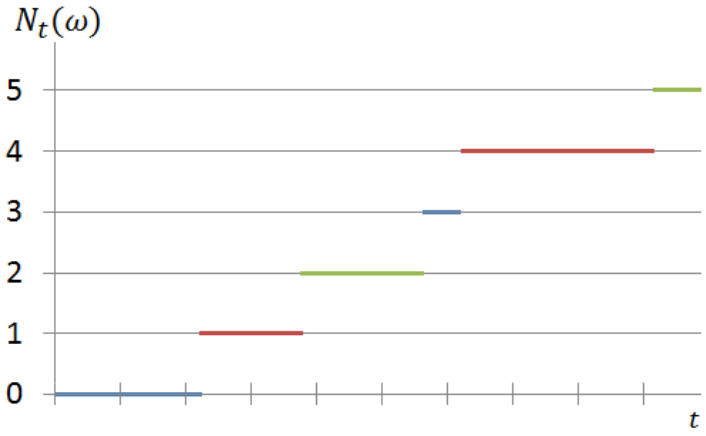


Figura 2.2: Posible resultado del proceso descrito con la variable N_t con $\omega = (2.2, 3.8, 5.6\ldots)$.

2.2.6. Distribuciones

La relación entre probabilidad y una *R.V.* X, en primera instancia, es designando al conjunto $\{\omega : X(\omega) \leq b\}$, o $\{X \leq b\}$, como un evento.

Analizaremos las distribuciones a causa de la correspondecia entre las *R.V.* y las medidas de probabilidad, definiendo las siguientes funciones

Definición 2.20. (Función de distribución acumulativa, CDF). *Sea X R.V. continua, se define el* **CDF** *de X, $F_X : \mathbb{R} \to [0, 1]$, como*

$$F_X(x) = P(X \leq x).$$

Propiedades:

- $0 \leq F_X(x) \leq 1$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to +\infty} F_X(x) = 1$.
- $x \leq y \Rightarrow F_X(x) \leq F_X(y)$.

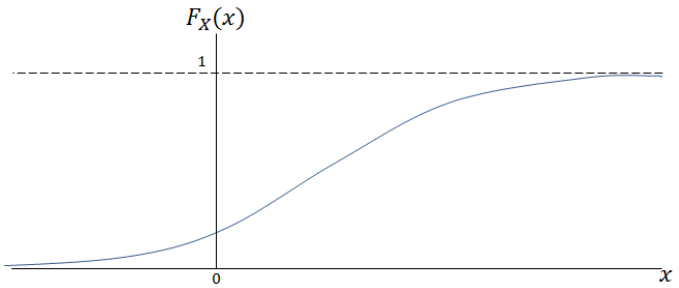


Figura 2.3: CDF.

Definición 2.21. (Función de densidad de probabilidad, PDF). *Sea X R.V continua, se define el* **PDF** *de X ,$f_X : \mathbb{R} \to [0, 1]$, como*

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Propiedades:

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{+\infty} f_X(x) = 1$.
- $F_X(x) = \int_{-\infty}^{x} f_X(u)du$.
- $P(b \leq X \leq a) = F_X(a) - F_X(b) = \int_{b}^{a} f_X(u)du$.

Definición 2.22. (Función de masa de probabilidad, PMF). *Sea X R.V. discreta, se define el* **PMF** *de X, $p_X : \Omega \to [0,1]$, como*

$$p_X(x) = P(X = x).$$

Propiedades:

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in X} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$ para todo $A \in \mathcal{F}$.

Definición 2.23. (Función de distribución acumulativa conjunta, J-CDF). *Sean X, Y R.V. continuas, se define el* **J-CDF** *de X, Y, $F_X : \mathbb{R} \times \mathbb{R} \to [0,1]$, como*

$$F_{XY}(x, y) = P(X \leq x, Y \leq Y).$$

Propiedades:

- $0 \leq F_{XY}(x, y) \leq 1$.
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$.
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$.

Sean X e Y R.V. independientes, entonces el J-CDF de X, Y cumpliría lo siguiente

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y).$$

Propiedades:

- $p_{XY}(x, y) = p_X(x)p_Y(y)$, si X e Y son R.V. discretas. Además, $p_{Y|X}(y|x) = p_Y(y)$ siempre que $p_X(x) \neq 0$.

- $f_{XY}(x,y) = f_X(x)f_Y(y)$, si X e Y son R.V. continuas. Además, $f_{Y|X}(y|x) = f_Y(y)$ siempre que $f_X(x) \neq 0$.

Definición 2.24. (Función de distribución acumulativa marginal, M-CDF). *Sean X, Y R.V. continuas, se definen los* **J-CDF**, $F_X : \mathbb{R} \times \mathbb{R} \to [0,1]$ *y* $F_Y : \mathbb{R} \times \mathbb{R} \to [0,1]$, *de X e Y respectivamente, como*

$$F_X(x) = \lim_{y\to\infty} F_{XY}(x,y) \ \wedge \ F_Y(y) = \lim_{x\to\infty} F_{XY}(x,y),$$

con las mismas propiedades que el CDF.

Definición 2.25. (Función de densidad de propabilidad conjunta, J-PDF). *Sean X, Y R.V. continuas, se define el* **J-PDF**, *de X e Y, como*

$$f_{XY}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}.$$

Propiedades:

- $f_{XY}(x,y) \geq 0$.
- $\displaystyle\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dydx = 1.$

Definición 2.26. (Función de densidad de propabilidad marginal, M-PDF). *Sean X, Y R.V., se definen f_X y f_Y, las* **J-PDF** *de X e Y respectivamente, como*

- *Si X, Y R.V. continuas*

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \ \wedge \ f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$

- *Si X es R.V. continua y Y es R.V. discreta*

$$f_X(x) = \sum_y f_{XY}(x,y).$$

con las mismas propiedades que el PDF.

Definición 2.27. (Función de masa de propabilidad conjunta, J-PMF). *Sean X, Y R.V. discretas, se define el* **J-PMF** *de X, Y, $p_X : Im(X) \times Im(Y) \to [0,1]$, como*

$$p_{XY}(x,y) = P(X = x, Y = y).$$

Propiedades:

- $0 \leq p_{XY}(x,y) \leq 1$.

- $\sum_x \sum_y p_{XY}(x,y) = 1.$

Definición 2.28. (Función de masa de propabilidad marginal, M-PMF). *Sean X, Y R.V. discretas, se definen p_X y p_Y, las* **J-PMF** *de X e Y respectivamente, como*

$$p_X(x) = \sum_y p_{X,Y}(x,y) \ \wedge \ p_Y(y) = \sum_x p_{X,Y}(x,y)$$

con las mismas propiedades que el PMF.

Definición 2.29. (Función de masa de probabilidad condicional, C-PMF). *Sean X, Y R.V. discretas, la C-PMF es definida como*

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)},$$

siempre que $p_X(x) \neq 0$.

Definición 2.30. (Función de densidad de probabilidad condicional, C-PDF). *Sean X, Y R.V. continuas, la C-PDF es definida como*

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Teorema 2.3. (Regla de Bayes). *Se establecen relaciones para las probabilidades condicionales como*

- *Si X, Y R.V. discretas,*

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y'} p_{X|Y}(x|y')p_Y(y')}$$

- *Si X, Y R.V. continuas,*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

2.2.7. Tendencias centrales y dispersión

Dado que es necesario la determinación de un valor típico como resultado de una serie de experimentos, se define la *esperanza* la cual determina este valor promedio. Además definimos cantidades relacionadas a las distribuciones con el objetivo de medir sus dispersiones.

Definición 2.31. (Esperanza). *Siendo X una R.V. definimos la media, o esperanza, como*

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x)dx, \text{ si } X \text{ es R.V. continua.} \\ E[X] &= \sum_x x p_X(x)dx, \text{ si } X \text{ es R.V. discreta.} \end{aligned}$$

De forma más general, sea $g : \mathbb{R} \to \mathbb{R}$ una función, luego $g(X)$ es también una R.V., cuya media es

$$\begin{aligned} E[f(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)dx, \text{ si } X \text{ es R.V. continua.} \\ E[f(X)] &= \sum_x g(x) p_X(x)dx, \text{ si } X \text{ es R.V. discreta.} \end{aligned}$$

Propiedades:

- $E[c] = c$ para toda $c \in \mathbb{R}$.
- $E[cg(X)] = cE[g(X)]$ para toda $c \in \mathbb{R}$.
- $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$.

La media nos es útil para determinar el valor esperado de un experimento aleatorio; sin embargo este valor no indica certeza por lo que se necesita tener en cuenta de que existe la posibilidad de que un valor resultante se desvíe de la expectativa; la varianza nos permite cuantificar aquello.

Definición 2.32. (Varianza 1^{era} definición). *Siendo X una R.V., se define la varianza como*

$$Var[X] = E[(X - E[X])^2]$$

De forma más general, sea $g : \mathbb{R} \to \mathbb{R}$ una función, luego $g(X)$ es también una R.V., cuya varianza es

$$Var[g(X)] = E[(g(X) - E[g(X)])^2]$$

La varianza nos establece el tipo de variabilidad de una variable aleatoria pues mide el riesgo de que un resultado no sea el valor esperado. En términos de distribución, nos mide la concentración que esta posee alrededor del valor esperado. En cierto aspectos se hace uso de la **desviación estándar** definida como la raíz cuadrada de la varianza.
Propiedades:

- $Var[cg(X)] = c^2 Var[g(X)]$ para toda $c \in \mathbb{R}$.

- $Var[g(X)] = E[g(X)^2] - E[g(X)]^2$.

Ahora toca interpretar la variabilidad entre varias R.V. conjuntas, para ello definiremos la *covarianza* de la siguiente forma

Definición 2.33. (Covarianza 1^{era} definición). *Sean X, Y R.V., se define su covarianza como:*

$$Cov(X,Y) = E[(X - E[X])(Y - E[X])].$$

De forma más general, sean $g : \mathbb{R} \to \mathbb{R}$ y $h : \mathbb{R} \to \mathbb{R}$ funciones reales, luego $g(X), h(Y)$ son R.V., cuya covarianza es

$$Cov[g(X), h(Y)] = E[(g(X) - E[g(X)])(h(Y)) - E[h(Y)])]$$

Como notamos, esta medida de variabilidad solo está definida para relacionar dos variables aleatorias. Para ello definiremos una matriz que relaciona simultáneamente las varianzas entres varias variables aleatorias y, que por su simetría, trae consigo importantes resultados.

Definición 2.34. (Matriz de Covarianza 1^{era} definición). *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias, entonces se define la matriz de covarianza como*

$$\Sigma = E[\mathbf{X}\mathbf{X}^T] = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix} \quad \textit{donde } \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

2.2.7.1. Estadística descriptiva

En estadística descriptiva estas medidas de variabilidad se presentan como

Definición 2.35. (Varianza 2^{da} definición). *Sean M datos de una variable X entonces la varianza se define como*

$$Var(X) = \frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{x})^2.$$

donde $\overline{x}$ es la media de los datos, x_i valores en cada dato respecto a X.

Definición 2.36. (Covarianza 2^{da} definición). *Sean M datos con dos variables X, Y, entonces la covarianza se define como*

$$Cov(X,Y) = \frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y}).$$

donde $\overline{x}, \overline{y}$ son la medias de los datos respectivamente a las variables X, Y.

Definición 2.37. (Matriz de Covarianza 2^{da} definición). *Sean M datos con n variables $X_1, \ldots, X_n$, entonces la matriz de covarianza se define como*

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix}.$$

	ϕ_1	ϕ_2	$\ldots$	ϕ_M
X_1	x_{11}	x_{12}	$\ldots$	x_{1M}
X_2	$\vdots$	$\ddots$		$\vdots$
$\vdots$				
X_n	x_{n1}	x_{n2}	$\ldots$	x_{nM}

Cuadro 2.1: M datos $\{\phi_1, \phi_2, \ldots\}$ con n variables; $\Phi_i = \{x_{1i}, \ldots, x_{ni}\}$.

Proposición 2.1. *Sea $\{\Phi_1, \ldots, \Phi_M\}$ un conjunto de M datos con n variables $X_1, \ldots, X_n$ con media igual a cero en cada una de ellas, entonces la matriz de covarianza viene determinada como*

$$\Sigma = \frac{1}{M} \begin{bmatrix} \Phi_1 & \ldots & \Phi_M \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \vdots \\ \Phi_M^T \end{bmatrix}$$

Demostración. Suponiendo que los M datos se presentan como en el Cuadro 2.1. Dado que la media es igual a cero en cada variable entonces tenemos que

$$cov(X_i, X_j) = \frac{1}{M} \sum_{k=1}^{M} x_{ik} y_{jk} \text{ para todo } i, j = 1, \ldots, n.$$

Ahora veamos lo siguiente, para concluir la prueba,

$$\frac{1}{M} \begin{bmatrix} \Phi_1 & \cdots & \Phi_M \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \vdots \\ \Phi_M^T \end{bmatrix} = \frac{1}{M} \begin{bmatrix} x_{11} & \ldots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nM} \end{bmatrix} \begin{bmatrix} x_{11} & \ldots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1M} & \ldots & x_{nM} \end{bmatrix}$$

$$= \frac{1}{M} \begin{bmatrix} \sum_{k=1}^{M} (x_{1k})^2 & \ldots & \sum_{k=1}^{M} x_{1k} x_{nk} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{M} x_{nk} x_{1k} & \ldots & \sum_{k=1}^{M} (x_{nk})^2 \end{bmatrix} = \begin{bmatrix} Cov(X_1, X_1) & \ldots & Cov(X_n, X_1) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \ldots & Cov(X_n, X_n) \end{bmatrix} = \Sigma$$

□

2.2.8. Distribución Gaussiana (Multivariable)

Las distribuciones Gaussianas son las más importante en su estudio sobre variables aleatorias pues nos dan grados de simétría ofreciéndonos visiones menos complejas que tratar con otro tipos de distribuciones.

Una distribución Gaussiana se define por tener el siguiente PDF sobre una variable aleatoria X,

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{2.5}$$

y se denota como

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{donde } \mu = E[X],\ Var[X] = \sigma^2.$$

Establezcamos lo siguiente $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ donde los $X_i's$ son una R.V.; veamos la distribución conjunta que determina el vector de variables aleatorias $\mathbf{X}$.

Proposición 2.2. *$X_1, \ldots, X_n$ variables aleatorias tales que $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ son independientes entonces, la distribución conjunta correspondiente viene determinada como*

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(\mathbf{x_i}) = \frac{1}{(2\pi)^{1/n} \prod_{i=1}^{n} \sigma_i} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\right) \tag{2.6}$$

Generalizando el resultado anterior, tenemos la siguiente proposición.

Proposición 2.3. *$X_1, \ldots, X_n$ variables aleatorias tales que $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ entonces resulta que tomar una distribución conjunta de estas variables nos genera una distribución denominada distribución Gaussiana multivariable,*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{1/n}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{2.7}$$

donde Σ es la matriz de covarianza, y μ es un vector con elementos μ_i.

Proposición 2.4. *Sean $\mathbf{x}_1, \mathbf{x}_2$ dos vectores variables con distribución Gaussiana multivariable, con representación*

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{21}^T \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

entonces,

$$\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}(\mu_1, A), && (2.8)\\ \mathbf{x}_2 &\sim \mathcal{N}(\mu_2, B), && (2.9)\\ \mathbf{x}_1|\mathbf{x}_2 &\sim \mathcal{N}(\mu_1 + \Sigma_{21}^T\Sigma_{22}^{-1}(\mathbf{y}-\mu_2), \Sigma_{11} - \Sigma_{21}^T\Sigma_{22}^{-1}\Sigma_{21}), && (2.10)\\ \mathbf{x}_2|\mathbf{x}_1 &\sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{y}-\mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^T). && (2.11) \end{aligned}$$

2.2.8.1. Intervalos de confianza

Una de las propiedades mas importantes es la del porcentaje de certeza (o confianza), sobre un intervalo, que nos relaciona la varianza alrededor de la media, esto es

$$P(\mu-\sigma < x \leq \mu+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) du = 0.682 = 68.2\,\%$$

Las descripciones se dan a continuación, y de forma análoga sobre otros intervalos de confianza (certeza), como se visualiza en el gráfico 2.4.

Propiedades:

- Existe 68.2 % de certeza entre 1 desviación estándar de la media.
- Existe 95 % de certeza entre 2 desviaciones estándar de la media.
- Existe 99.7 % de certeza entre 3 desviaciones estándar de la media.

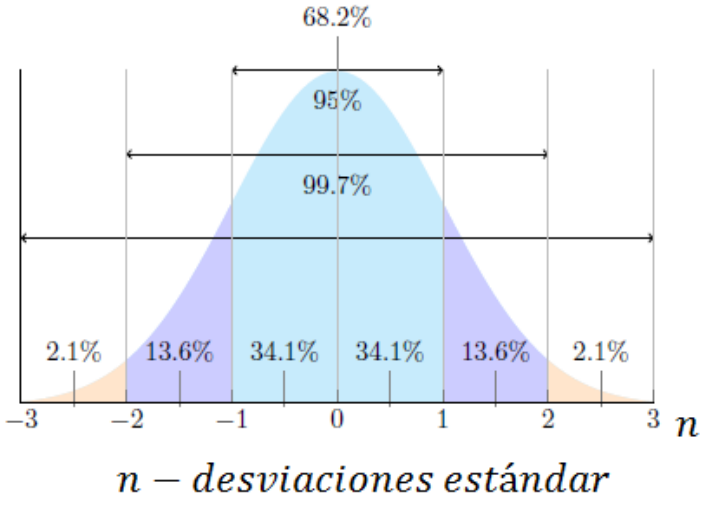


Figura 2.4: Porcentaje de certeza entre desviaciones estándar de la media.

2.2.9. Estadística Inferencial

La inferencia estadística, nos proporciona enfoques de inducción sobre datos observados y comprende estimación de parámetros que es parte de lo que necesitamos enfatizar.

2.2.9.1. Principio de Verosimilitud

Como se vio anteriormente, la probabilidad condicional nos proporciona el resultado probabilístico de que ocurra un evento A, habiendo ya ocurrido un evento B como $P(A|B)$; el análisis es análogo para el C-PDF de dos R.V. continuas, por lo que si deseamos determinarla para $Y = y$ sabiendo de un resultado eventual $\Theta = \theta$ (denominado parámetro), se determinaría el correspondiente $f_{Y|\Theta}(y|\theta)$.
La **verosimilitud**, o principio de verosimilitud, nos representa una intuición con respecto a $\Theta = \theta$ de ser un parámetro desconocido, habiéndose ya obtenido $Y = y$, por lo que, estadísticamente, está relacionada con la probabilidad condicional en forma inversa, definiendo la *función de verosimilitud* como
(Evitando abuso de notación $f_{Y|\Theta} = f$)

$$\mathcal{L}(\theta|y) \quad = \quad f(y|\theta).$$

En forma general, se tiene la siguiente definición

Definición 2.38. (Función de verosimilitud). *Siendo una muestra aleatoria $Y_1, \ldots, Y_n$, se define la función de verosimilitud como*

$$\mathcal{L}(\theta|\mathbf{y}) = f(\mathbf{y}|\theta), \qquad \textit{donde } \mathbf{y} = y_1, \ldots, y_n.\ (y_i \in Y_i)$$

De ser los $Y_i's$ independientes, entonces

$$\mathcal{L}(\theta|\mathbf{y}) \quad = \quad \prod_{i=1}^{n} f(\mathbf{y}_i|\theta). \tag{2.12}$$

2.2.9.2. Estimador de Máxima Verosimilitud (E.V.M.)

Un importante uso de esta función de verosimilitud es que logra establecer el *estimador de máxima verosimilitud* que es el parámetro más próximo que define a una población en base a una toma de muestras $Y_1 = y_1, \ldots, Y_n = y_n$ de una población que son regidas por una distribución $f(\cdot|\theta)$ intuitivamente.

El **estimador de máxima verosimilitud** (E.V.M) se determina como:

$$\tilde{\theta} = \max_{\theta}\{\mathcal{L}(\theta|\mathbf{y})\}. \tag{2.13}$$

Debido a que la función $\ln : \mathbb{R}_+ \to \mathbb{R}$ es creciente, en muchas ocasiones la estimación de máxima verosimilitud corresponde a determinar

$$\tilde{\theta} = \max_{\theta}\{l(\theta|\mathbf{y})\} \tag{2.14}$$

donde $l(\theta|\mathbf{y}) = \ln(\mathcal{L}(\theta|\mathbf{y}))$.

2.2.9.3. Estimador de Máximo a Posteriori (M.A.P.)

De existir la correspondiente distribución para el parámetro θ, entonces el análisis bayesiano implica que la *regla de Bayes* tomaría la siguiente forma

$$\pi(\theta|\mathbf{y}) \quad = \quad \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} = \frac{\mathcal{L}(\theta|\mathbf{y})\pi(\theta)}{\displaystyle\int_{-\infty}^{+\infty} f(\mathbf{y}|\theta')\pi(\theta')d\theta'}.$$

donde $\pi(\theta)$ se conoce como **distribución a priori**, ya que representa información no condicionda por los datos; y $\pi(\theta|\mathbf{y})$ se conoce como **distribución a posteriori**, ya que representa información habiéndose ya optado por un conjunto de datos $y_1, \ldots, y_n$.

Una mejor estimación del parámetro θ, se denomina como **estimador de máximo a posteriori**, y corresponde a determinar

$$\tilde{\theta} = \max_{\theta}\{\mathcal{L}(\theta|\mathbf{y})\pi(\theta)\}. \tag{2.15}$$

2.3. Optimización

Los conceptos de optimización, realizada en esta sección es para enmarcar una técnica de minimización de funciones. En primera instancia, un problema de minimización se denota como

$$\mathcal{O}\begin{cases} x_{\text{mín}} = \min\limits_{x}\{F(x)\} \\ x \in Dom(F) \subset \mathbb{R}^n. \end{cases} \tag{2.16}$$

donde a la función F, comúnmente se la denomina *Función Costo*.

2.3.1. Convexidad

Definición 2.39. Conjunto Convexo. *Un conjunto $C \subset \mathbb{R}^n$ es convexo si para todo $x, y \in C$ y $\lambda \in [0,1]$ se tiene que $(1-\lambda)x + \lambda y \in C$.*

Definición 2.40. Cápsula Convexa. *Sea $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^n$ subconjunto finito, la cápsula convexa de X, denotado por $conv(X)$ se define como el conjunto de todas las combinaciones convexas de $x_1, \ldots, x_m$, es decir*

$$conv(X) = \left\{ \sum_i \lambda_i x_i : \quad x_i \in X, \lambda_i \geq 0, \sum_i \lambda_i = 1 \right\}.$$

Definición 2.41. Función convexa. *Sea $U \subset \mathbb{R}^n$ convexo, se dice que una función $f : U \subset \mathbb{R}^n \to \mathbb{R}$ es convexa si para todo $x, y \in U$ y $\lambda \in [0, 1]$ se tiene que*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

2.3.2. Descenso del Gradiente

El descenso de gradiente es una de tantas técnicas de minimización de funciones costo, en particular analizaremos esta técnica para funciones convexas.
La técnica presenta el siguiente algoritmo

Algoritmo 1: Descenso de Gradiente

Datos: $\mathbf{x}^0 \in \mathbb{R}^n$ (punto inicial), γ (tolerancia).
1 $t \leftarrow 0$;`/* Media de los datos. */`
2 **mientras** $\|\nabla F(\mathbf{x}^0)\| \geq \gamma$ **hacer**
3 $\quad \mathbf{x}^{t+1} \leftarrow \mathbf{x}^t - \alpha \nabla F(\mathbf{x}^t)$; $t \leftarrow t + 1$
4 **fin mientras**
5 **devolver** $\mathbf{x}^t$

donde x_0 es el punto de partida hacia la obtención del mínimo.

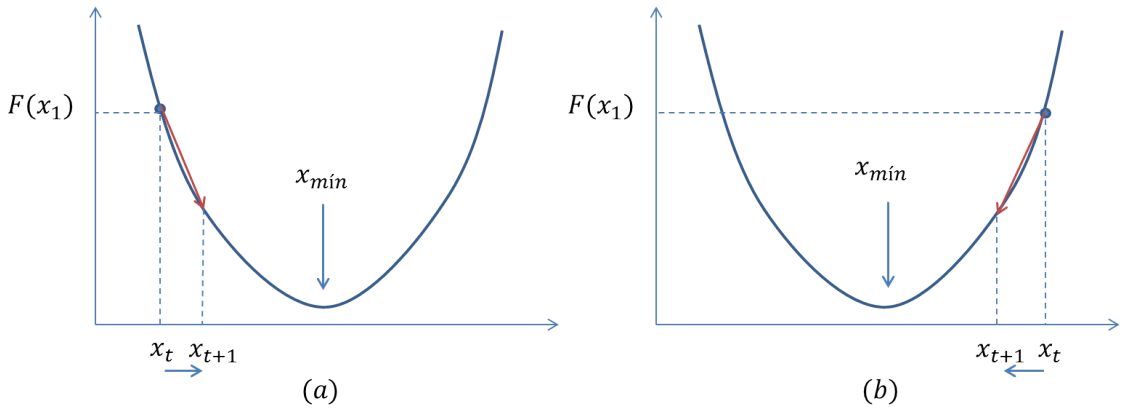


Figura 2.5: Visualizacion del proceso llevado a cabo por la técnica de descenso de gradiente, en base a una función costo $F : U \subset \mathbb{R} \to \mathbb{R}$.

Además, hay que tener en cuenta que si optamos por un γ en nuestro algoritmo *Descenso de gradiente* (línea 3 - Algoritmo 1), de modo que sea pequeño, entonces hará de

nuestra convergencia un proceso lento; sin embargo, de optar por un γ muy grande, el proceso podría diverger y por tanto no establecer el punto mínimo. (ver Figura 2.6)

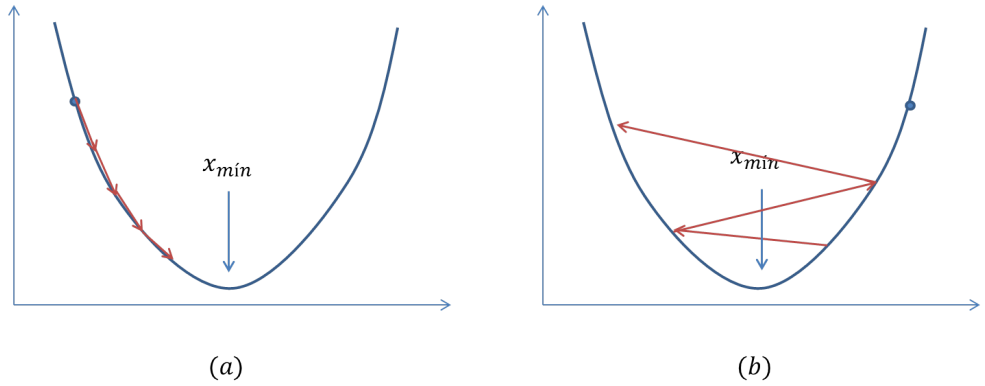


Figura 2.6: (a) Visualización de un valor γ apropiado; y (b) uno inapropiado que haga el proceso divergente.

Capítulo 3

Regresión Lineal

Las técnicas de regresión lineal estiman, mediante rectas, las relaciones entre las carácterísticas propias de los datos y valores numéricos de salida dependientes de estas características. El estudio desarrollado por estas técnicas son parte fundamental en el campo de la predicción, que a su vez está incorporada dentro del aprendizaje supervisado. El enfoque dado en este capítulo es favoreciéndo a un análisis bayesiano que nos interesa en el siguiente capítulo.
El desarrollo de las técnicas vistas serán asumiendo n datos de entrenamiento que diseñen el modelo.

3.1. Regresión Lineal Simple

La regresión lineal simple se basa en técnicas para datos unidimensionales, es decir, con solo una característica (variable) que generan valores de salida numéricos por cada dato. Esta técnica supondrá contar con un solo valor de salida, pues de tenerse más se desarrollan modelos de forma similar para cada uno de ellos.
Por tanto, el modelo tiene la siguiente forma

$$Y = h_\theta(x) + \epsilon \quad \text{con } h_\theta(x) = \theta_0 + \theta_1(x) \tag{3.1}$$

donde θ_0, θ_1 son los coeficientes de regresión. Asumiendo $\epsilon \sim \mathcal{N}(0, \sigma^2)$, esto es

$$E[\epsilon] = 0 \quad \text{y} \quad Var[\epsilon] = \sigma^2. \tag{3.2}$$

La denominación para ϵ es *Gaussian noise*, o ruido Gaussiano.

Se representarán los datos de entrenamiento como tuplas (x_i, y_i) para $i = 1, \ldots, n$, donde x_i sean los valores de su primera variable y y_i los valores de salida.

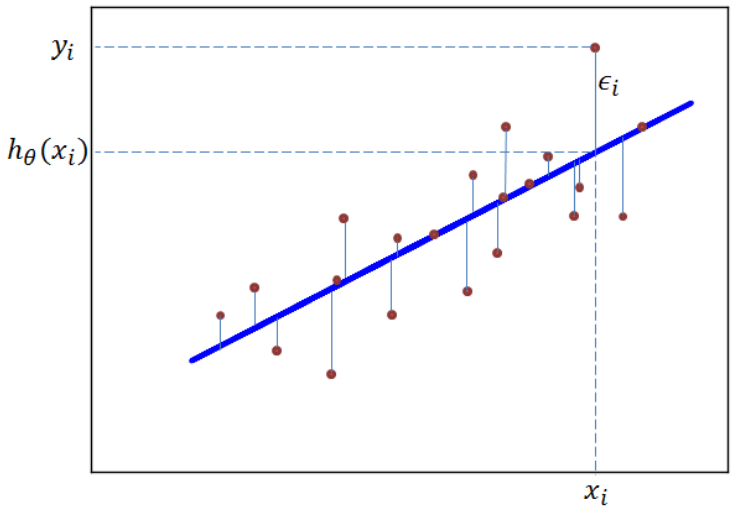


Figura 3.1: Regresión lineal simple (Análisis).

Sean Y, X, Θ variables aleatorias, representando los valores de salida, características y coeficientes de regresión, y usando la probabilidad condicional de Y sobre $X = x$ y $\Theta = \theta = (\theta_0, \theta_1)$, denominado *parámetro*, tenemos lo siguiente

$$E[Y|x,\theta] = E[h_\theta(x) + \epsilon] = h_\theta(x) + E[\epsilon] = h_\theta(x). \tag{3.3}$$

Además,

$$\begin{aligned} Var[Y|x,\theta] &= E[(Y|x,\theta)^2] - E[(Y|x,\theta)]^2 \\ &= E[(h_\theta(x) + \epsilon)^2] - h_\theta(x)^2 \\ &= (h_\theta^2(x) + 2h_\theta(x) \underbrace{E[\epsilon]}_{=0} + \underbrace{E[\epsilon^2]}_{=\sigma^2}) - h_\theta^2(x) \\ &= \sigma^2 \end{aligned} \tag{3.4}$$

Indicando que $Y|x,\theta \sim \mathcal{N}(h_\theta(x), \sigma^2)$. Ahora asumiendo que para los $x_i's$, la variables $Y_i's$ son independientes; usando la regla de chain, dado que $Y_i|x,\theta \sim \mathcal{N}(h_\theta(x_i), \sigma^2)$.

$$\begin{aligned} f(\mathbf{y}|x_1,\ldots,x_n,\theta) &= \prod_{i=1}^{n} f(y_i|x_i,\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{|\mathbf{y} - X^T\theta|^2}{2\sigma^2}\right) \end{aligned} \tag{3.5}$$

donde $X = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$ y $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$.

Por el *principio de verosimilitud*,

$$\mathcal{L}(\theta|X,\mathbf{y}) = f(\mathbf{y}|x_1,\ldots,x_n,\theta) \tag{3.6}$$

estimamos el parámetro θ establecido por (2.14) la *máxima verosimilitud*

$$\tilde{\theta} = \arg\max_{\theta}\{l(\theta|X,\mathbf{y})\}, \quad \text{donde } l(\theta|X,\mathbf{y}) = \ln(\mathcal{L}(\theta|X,\mathbf{y})). \tag{3.7}$$

Esto nos conduce a determinar el parámetro que maximice la siguiente función

$$l(\theta|X,\mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2. \tag{3.8}$$

3.1.1. Función Costo (Minimización)

Lo anteriormente establecido nos corresponde a minimizar

$$\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 \quad = \quad \sum_{i=1}^{n}(y_i - \theta_0 + \theta_1 x_i)^2 \tag{3.9}$$

o equivalentemente el error medio cuadrático.

$$\tilde{\theta} = \text{mín}\left\{F(\theta)\right\}, \quad \text{donde } F(\theta) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2. \tag{3.10}$$

donde F la denominaremos como *función costo*, que es el término usual en problemas de optimización.

Siendo la función costo, una función convexa, entonces el uso de la técnica de descenso de gradiente la minimiza hacia un resultado global, teniendo en cuenta que no es la única técnica que se puede optar para ello. (sección 2.3.2. - Algoritmo 1)

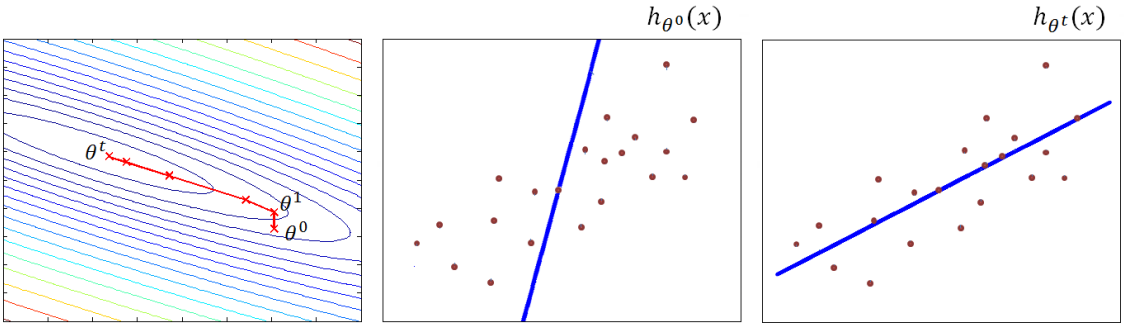


Figura 3.2: Descenso de Gradiente para técnicas de regresión lineal simple.

3.2. Regresión lineal múltiple

La técnica de regresión lineal múltiple mantiene un modelo análogo al de regresión lineal simple con la diferencia de que se aplica para datos multidimensionales. Al igual que en regresión lineal simple, suponemos que existe solo un valor de salida por cada dato.
El modelo toma la siguiente forma, suponiendo n datos $k-dimensionales$ (datos de entrenamiento),

$$Y = h_\theta(\mathbf{x}) + \epsilon, \qquad \text{con } h_\theta(\mathbf{x}) = \mathbf{x}^T\theta, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{bmatrix} \tag{3.11}$$

donde θ se denomina *parámetro*, cuyos elementos son los *coeficientes de regresión*, y suponiendo que $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (ϵ, ruido Gaussiano) no correlacionados.

Representando los datos de entrenamiento como tuplas $(\mathbf{x}_i, y_i)$ para $i = 1 \ldots, n$, donde $\mathbf{x}_i = x_{i1}, \ldots, x_{ik}$ contienen los valores en cada una de sus dimensiones, y y_i sea el valor de salida. Denotemos matricialmente los vectores de características de cada dato de entrenamiento, valores de salida y ruidos Gaussianos como

$$X = \begin{bmatrix} 1 & \ldots & 1 \\ x_{11} & & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \ldots & x_{kn} \end{bmatrix}, \quad \mathbf{y} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3.12}$$

entonces tenemos una representación matricial del modelo de la siguiente forma

$$\mathbf{y} = X^T\theta + \varepsilon, \text{ con } \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{n\times n}). \tag{3.13}$$

Ahora a causa de la independencía de ϵ_i, tenemos por Proposición 2.2

$$f(\varepsilon) = -\frac{1}{(2\pi)^{1/n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon\right) \tag{3.14}$$

por (3.13), $\varepsilon = \mathbf{y} - X^T\theta$, redefinimos (3.14) como

$$\mathcal{L}(\theta|X, \mathbf{y}) = f(\mathbf{y}|X, \theta) = -\frac{1}{(2\pi)^{1/n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta)\right). \tag{3.15}$$

y por tanto podemos determinar los E.V.M. que estime nuestro parámetro θ, entonces el

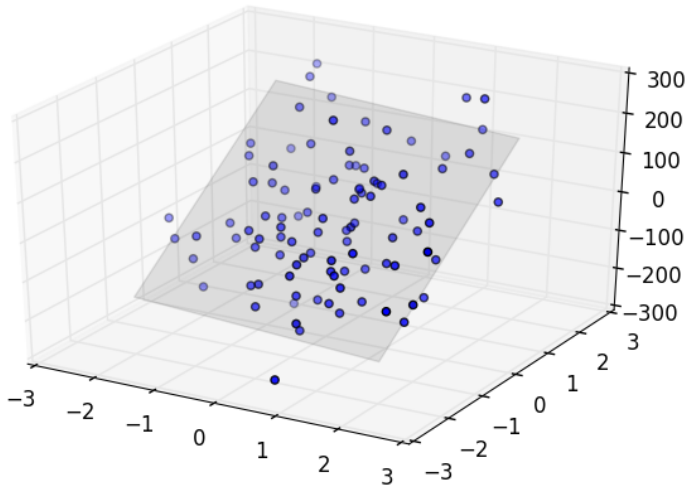


Figura 3.3: Regresión lineal múltiple para datos bidimensionales (Visualización). Representación del modelo: $y = h_\theta \mathbf{x} + \epsilon = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon$.

problema a tratar es maximizar lo siguiente,

$$l(\theta|X, \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta). \tag{3.16}$$

3.2.1. Función Costo (Minimización)

El resultado anterior nos conlleva a un problema de minimización de errores medios cuadrátricos, análogo para regresión lineal simple, dada por

$$F(\theta) = \frac{1}{n}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k}\theta x_j\right). \tag{3.17}$$

Siendo esta función costo una función convexa, entonces no presenta mínimos locales y por tanto el uso de la técnica de descenso de gradiente permite minimizarla. (sección 2.3.2. - Algoritmo 1)

3.3. Análisis Bayesiano

Nuestro análisis para las técnicas de regresión lineal simple y múltiple, nos permitió considerar la siguiente función de verosimilitud

$$\mathcal{L}(\theta|X, \mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\|\mathbf{y} - X^T\theta\|^2}{2\sigma^2}\right) = \mathcal{N}(X^T\theta, \sigma^2\mathbb{I}_{n\times n}). \tag{3.18}$$

Ahora por la reformulación de la regla de Bayes (Teorema 2.3) tenemos lo siguiente

$$\pi(\theta|\mathbf{y}, X) = \frac{\mathcal{L}(\theta|X, \mathbf{y})\pi(\theta)}{\int_{-\infty}^{+\infty} f(\mathbf{y}|X, \theta')\pi(\theta')d\theta'} \tag{3.19}$$

Notamos que , por *marginalización condicionada*, la distribución marginal

$$f(\mathbf{y}|X) = \int_{-\infty}^{+\infty} f(\mathbf{y}|X, \theta')\pi(\theta')d\theta' \tag{3.20}$$

es independiente con el parámetro θ.

La inferencia bayesiana sobre nuestro modelo se basa en analizar la función a posteriori, asumiendo una distribución a priori $\pi(\theta) \sim \mathcal{N}(\mathbf{0}, \widehat{\Sigma})$, y a causa de (3.20)

$$\pi(\theta|X, \mathbf{y}) \propto \mathcal{L}(\theta|X, \mathbf{y})\pi(\theta) \tag{3.21}$$

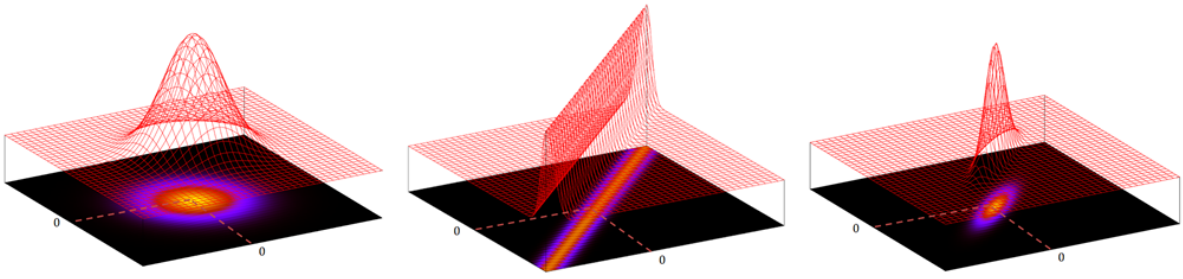
$$\pi(\theta|X, \mathbf{y}) \propto \exp\left(-\frac{\|\mathbf{y} - X^T\theta\|^2}{2\sigma^2}\right) \exp\left(-\frac{\theta^T\widehat{\Sigma}^{-1}\theta}{2}\right)$$

$$\pi(\theta|X, \mathbf{y}) \propto \exp\left(\frac{1}{2}(\theta - \overline{\theta})^T A^{-1}(\theta - \overline{\theta})\right) \tag{3.22}$$

$$\tag{3.23}$$

por tanto

$$\pi(\theta|\mathbf{y}, X) = \mathcal{N}(\overline{\theta}, A^{-1}) \quad \text{donde} \quad \begin{cases} A = \sigma^{-2}XX^T + \widehat{\Sigma}^{-1} \\ \overline{\theta} = \sigma^{-2}A^{-1}X\mathbf{y}. \end{cases} \tag{3.24}$$



$\pi(\theta) \sim N(0, \mathbb{I}_{2\times 2})$ $f(y|X,\theta) \sim N(X^T\theta, \sigma^2\mathbb{I}_{2\times 2})$ $\pi(\theta|X,y) \sim N(\theta, A^{-1})$

(a) (b) (c)

Figura 3.4: Sea el modelo de regresión $y = \theta_0 + \theta_1 x + \epsilon$, (a) representa la distribución a priori $\pi(\theta) \sim \mathcal{N}(0, \mathbb{I}_{2\times 2})$; (b) la densidad de probabilidad de las observaciones dado los parámetros $f(y|X, \theta) \sim \mathcal{N}(X^T\theta, \sigma^2\mathbb{I}_{n\times n})$; (c) representa la distribución a posteriori $\pi(\theta|\mathbf{y}, X) \sim \mathcal{N}(\overline{\theta}, A^{-1})$.

Capítulo 4

Procesos Gaussianos

En este capítulo, el análisis de los Procesos Gaussianos está enfocado hacia técnicas de regresión con el objetivo de diseñar un algoritmo de aprendizaje automatizado.

Básicamente, los procesos gaussianos (GP) manejan su utilidad como distribuciones sobre funciones, formalmente denominamos a un proceso gaussiano como un proceso estocástico continuo $\{X_i : \ i \in \mathbb{R}\}$ donde cada subconjunto finito de R.V.'s presentan una distribución gaussiana multivariable.

4.1. Motivación

Suponiendo que tenemos n datos $k - dimensionales$ representados como $(\mathbf{x}_i, y_i)$ para todo $i = 1, \ldots, n$, donde $\mathbf{x}_i \in \mathbb{R}^k$ es el vector de valores de cada dimensión en los datos y $y_i \in \mathbb{R}$ los valores de salida correspondientes; claramente podemos modelar una técnica de regresión lineal que ajuste los datos (datos de entrenamiento) para la predicción de un nuevo valor de salida para un dato con vector de características $\mathbf{x}^*$. Sin embargo los procesos gaussianos , modelan una técnica basándose en un análisis bayesiano análogo al estudiado para regresiones lineales en el capítulo anterior.

La representación del modelo es

$$y = h(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.1}$$

y un proceso gaussiano (**GP**), cuyo técnica se basa en distribuciones sobre funciones, representando datos como una muestra de una distribución multivariada, se denota como

$$h(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{4.2}$$

donde $m(x)$ es la función promedio, y k una función de covarianza sobre funciones.

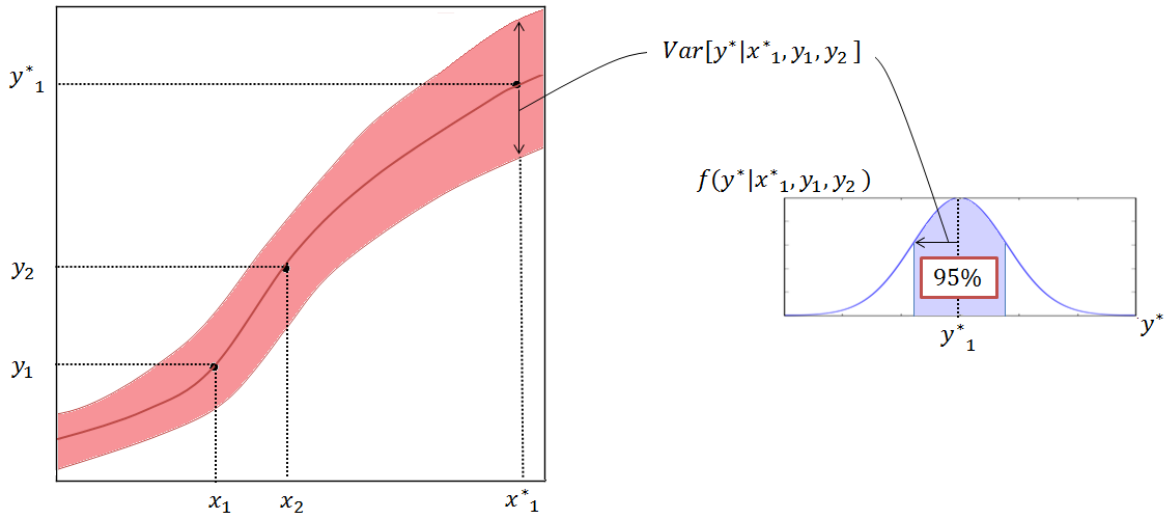


Figura 4.1: Teniendo dos datos de entrenamiento unidimensionales, el proceso gaussiano para regresión nos predice $y_1^* = E[y^*|y_1, y_2]$, para cierto valor característico x_1^*, que se encuentra dentro de un intervalo de confianza del 95 % proporcionada por 2 varianzas alrededor de y_1^*.

La técnica que se estudiará se denomina *técnica de regresión por procesos gaussianos* (**GPR**) y nos establece un modelo que teniendo un vector de características $\mathbf{x}^*$ para un dato nos predice su valor de salida y^* siendo este el valor esperado de una distribución condicional y^* con respecto a demás valores de salida ya establecidos por nuestros datos de entrenamiento, esto es

$$f(y^*|\mathbf{x}^*, y_1, \ldots, y_n) \tag{4.3}$$

, además de ortorgar un intervalo de confianza medida por la varianza de dicha distribución, como se visualiza en Figura 4.1.

Como notamos esta regresión no es lineal, pero lo será en un espacio de mayor dimensión, que es el espacio en donde estos procesos fundamentan su análisis, lo cual será visto en la sección posterior.

Suponiendo haberse hecho el correspondiente análisis y diseñado el método, se llegará a establecer una relación directa de este con la matriz de covarianza de los datos regidos bajo ciertas funciones, entendiendo que esta matriz no la de Definición 2.37; resumiendo explicaciones y evitando complejidades, esta matriz de covarianza para n datos unidimensionales, con varianza σ^2 (4.1), toma la siguiente forma

$$\tilde{\Sigma} = \begin{bmatrix} \kappa(x_1, x_1) & \ldots & \kappa(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n, x_1) & \ldots & \kappa(x_n, x_n) \end{bmatrix} \quad \text{donde } \kappa(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2}\right) \tag{4.4}$$

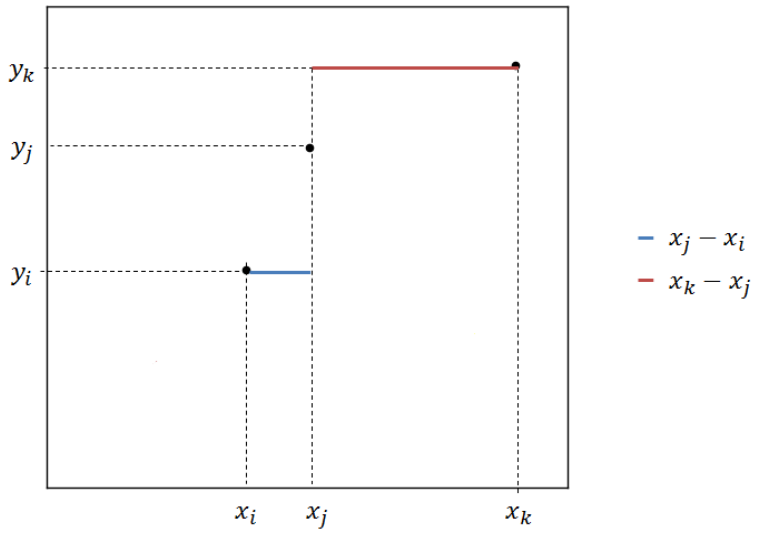


Figura 4.2: Se visualizan 3 datos unidimensionales, y se logra notar una menor distancia entre x_i a x_j que de x_k a x_j, por lo que la función de covarianza exponencial cuadrática, $\kappa(\cdot,\cdot)$ (4.4) nos asegura una relación más "fuerte" $x_i \leftrightarrow x_j$.

y la matriz de covarianza ampliada por un nuevo valor de características x^* tendría la siguiente representación

$$\begin{bmatrix} & & & \kappa(x_1, x^*) \\ & \tilde{\Sigma} & & \vdots \\ & & & \kappa(x_n, x^*) \\ \kappa(x^*, x_1) & \dots & \kappa(x^*, x_n) & \kappa(x^*, x^*) \end{bmatrix} = \begin{bmatrix} \tilde{\Sigma} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix} \tag{4.5}$$

La función

$$\kappa(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}^+ \tag{4.6}$$

es una *función Kernel*, la cual será descrita en el transcurso del capítulo, y para este caso en particular (4.4), se denomina *función de covarianza exponencial cuadrática*, la cual estima relaciones mas altas entre datos más cercanos, es decir que según como se establece esta función si tenemos tres valores de características x_i, x_j, x_k, como se visualiza en Figura 4.2, entonces

$$\text{si } |x_j - x_i| < |x_j - x_k| \quad \Rightarrow \quad \kappa(x_j, x_i) > \kappa(x_j, x_k). \tag{4.7}$$

y además se observa que,

$$\text{si } \lim_{x_i \to x_j} \kappa(x_j, x_i) = \kappa(x_j, x_j) = 1 \tag{4.8}$$

$$\text{si } \lim_{x_i \to \infty} \kappa(x_j, x_i) = 0. \tag{4.9}$$

Ahora, en base al concepto informal del proceso gaussiano (4.2), el vector de valores de salida mantendría una distribución multivariada, con la adición del valor de salida y^* que se desea predecir nos llevaría, usando (4.5), a lo siguiente

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Sigma} + \sigma^2\mathbb{I} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix}\right) \tag{4.10}$$

Sin embargo para la técnica GPR, necesitamos establecer la función de distribución condicional $f(y^*|\mathbf{y})$, la cual, sin mayor explicaciones, presenta la siguiente distribución

$$y^*|\mathbf{x}^*, \mathbf{y} \sim \mathcal{N}\left(\Sigma^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y}, \Sigma^{**} - \Sigma^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\Sigma^{*T}\right) \tag{4.11}$$

donde $E[y^*|\mathbf{y}] = \Sigma^*\tilde{\Sigma}^{-1}\mathbf{y}$ y $Var[y^*|\mathbf{y}] = \Sigma^{**} - \Sigma^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\Sigma^{*T}$, siendo estos valores los que determinen el algoritmo de aprendizaje que nos interesa en este proyecto.

4.2. Análisis Constructivos

Para hacer uso correspondiente de lo que entendemos por Procesos Gaussianos, optaremos primero por considerar ciertos análisis que nos lleven al objetivo de la construcción de la técnica **GPR**. En la sección anterior, se informó el proceso que tomaría un algoritmo de aprendizaje que use la técnica de regresión por procesos Gaussianos (GPR); sin embargo, se optó por evitar explicaciones con el fin de considerar ideas básicas de estadística y probabilidad. El enfoque ahora será sumergir ideas más complejas que fundamenten la técnica.

4.2.1. Primer Análisis (Distribución predictiva)

Planteemos una primera visión para el análisis de regresión lineal, considerando y^* el valor predictivo que logra establecer el modelo (3.11), la cual viene condicionada a una características o vector de características $\mathbf{x}^*$ que se pone a prueba, esto es $y^* = h_\theta(\mathbf{x}^*)$, pero existen múltiples parámetros que logren determinar un valor de salida; por tanto hacer predicciones sobre dicho valor nos corresponde definir una *distribución predictiva* en base al promedio de todos los posibles parámetros; esto se logra establecer con el método de marginalización condicionada

$$f(y^*|\mathbf{x}^*, \mathbf{y}) = \int f(y^*|x^*, \theta')\pi(\theta'|X, \mathbf{y})d\theta' \tag{4.12}$$

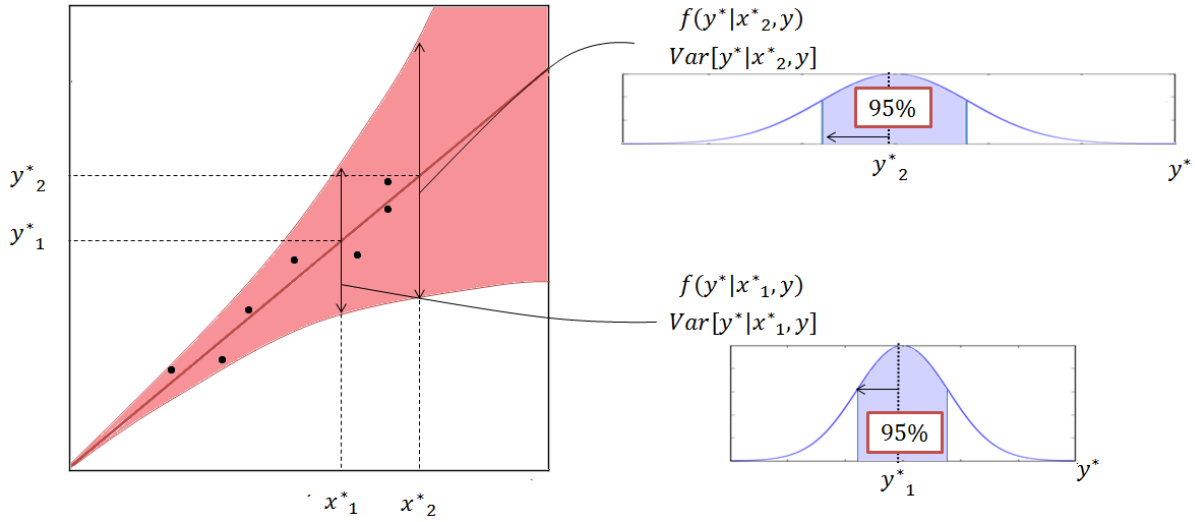


Figura 4.3: Teniendo datos de entrenamiento unidimensionales, se visualiza el incremento de varianza de las distribuciones predictivas conforme el valor característico sea más grande. Por tanto la incertidumbre sobre nuestro valor predicho crece. En particular se visualiza el intervalo de confianza del 95 % de certeza que nos proporciona 2 varianzas alrededor de la media.

La marginalización esta haciendo uso de (3.19), $\pi(\theta|X, \mathbf{y}) = \mathcal{N}\left(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}\right)$, luego tenemos que

$$f(y^*|\mathbf{x}^*, \mathbf{y}) = \mathcal{N}\left(\sigma^{-2}\mathbf{x}^{*T}A^{-1}X\mathbf{y}, \mathbf{x}^{*T}A^{-1}\mathbf{x}^*\right). \tag{4.13}$$

Sea Y^* la variable aleatoria que corresponde a los valores de salida predichos. Como observamos, la *distribución predictiva* presenta una distribución Gaussiana con varianza cuadrática $Var[Y^*|\mathbf{x}^*, \mathbf{y}] = \mathbf{x}^{*T}A^{-1}\mathbf{x}^*$, esto nos da a entender que a mayor magnitud del vector de carácterísticas $\mathbf{x}^*$, del dato para el cual se desea predecir su valor de salida, la varianza de la distribución predictiva aumenta y por lo tanto la incertidumbre.

Habiéndose hecho este análisis en base al modelo de regresión lineal, e intuir que

$$Var[Y^*|\mathbf{x}^*, \mathbf{y}] \to \infty \quad \text{si } \|\mathbf{x}^*\| \to \infty. \tag{4.14}$$

entonces el estudio para dichas distribuciones predictivas no satisface el grado de certeza que uno desease considerar, es ahí cuando partimos por realizar un análisis constructivo que nos mantenga un intervalo de certeza menos amplio.

Las comparaciones entre las Figuras 4.1 y 4.3, visualiza la diferencia sobre estas distribuciones predictivas con el uso de procesos gaussianos y sin su uso.

4.2.2. Segundo Análisis (Extensión de la Dimensionalidad)

En regresión lineal se define un modelo que logra establecer básicamente un subespacio $k-dimensional$, para n datos de entrenamiento $k-dimensionales$; sin embargo, podemos aplicar el modelo extendiendo la dimensionalidad de k a K $(k < K)$ bajo los mismos datos $k - dimensionales$.

Matemáticamente, el modelo a tratar sería el siguiente

$$\begin{gathered} y = h_\theta(\mathbf{x}) + \epsilon, \text{ donde } h_\theta(\mathbf{x}) = \phi(\mathbf{x})^T\theta \\ \text{y } \ \phi : \mathbb{R}^k \to \mathbb{R}^K (k < K) \end{gathered} \tag{4.15}$$

con $\epsilon \sim \mathcal{N}(0, \sigma^2)$, siendo $\mathbf{x} = (x_1, \ldots, x_k)$ el vector de características para un dato $k - dimensional$ y θ el parámetro que, a diferencia del modelo usual de regresión lineal, sería $K - dimensional$

Como notamos, el modelo se representa con una función ϕ que mapea las características de un dato hacia un espacio de mayor dimensión, y en ese nuevo espacio de característica es donde el modelo lineal anteriormente estudiado toma una forma similar, como se visualiza en el ejemplo de la Figura 4.4.

La función ϕ, por como está definida podemos representarla como

$$\phi(\mathbf{x}) = \begin{bmatrix} \phi_1(\mathbf{x} \\ \vdots \\ \phi_K(\mathbf{x}) \end{bmatrix}. \tag{4.16}$$

Por tanto, de nuestra representación de vectores de características, para n datos de entrenamiento $(\mathbf{x}_i, y_i)$ para todo $i = 1, \ldots, n$ $k - dimensionales$, mediante una matriz X, vista con anterioridad, la función mapea dicha matriz como sigue a continuación

$$X = \begin{array}{c} \begin{matrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{matrix} \\ \begin{bmatrix} 1 & \ldots & 1 \\ x_{11} & \ldots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \ldots & x_{kn} \end{bmatrix} \end{array} \xrightarrow{\phi} \begin{array}{c} \begin{matrix} \phi(\mathbf{x_1}) & \cdots & \phi(\mathbf{x_n}) \end{matrix} \\ \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_1(\mathbf{x_n}) \\ \phi_2(\mathbf{x}_1) & \ldots & \phi_2(\mathbf{x_n}) \\ \vdots & \ddots & \vdots \\ \phi_K(\mathbf{x}_1) & \ldots & \phi_K(\mathbf{x_n}) \end{bmatrix} \end{array} = \Phi(X). \tag{4.17}$$

denotando como $\Phi(X)$ a la matriz de los vectores de características, para los datos de entrenamiento, bajo las funciones ϕ.

El análisis bayesiano para este nuevo modelo, representaría una variación con respecto

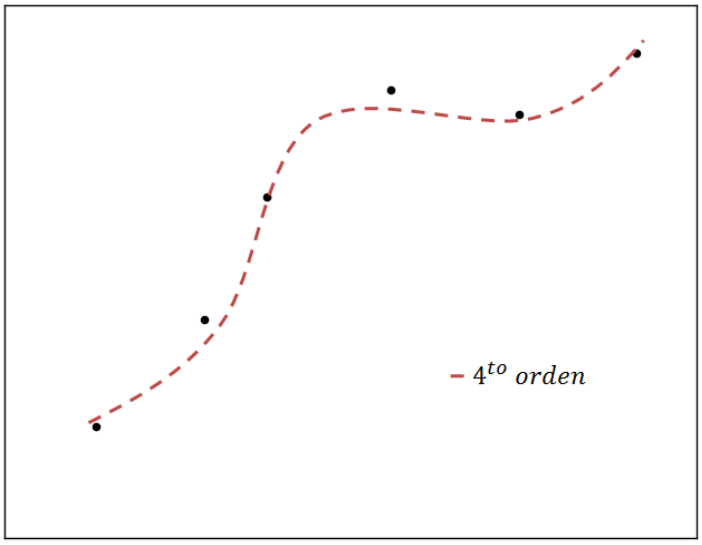


Figura 4.4: datos unidimensionales, siendo x el valor de la caraterística y $\phi(x) = (1, x, x^2, x^3, x^4)$ entonces el modelo se representaría como $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \epsilon$, el cual se denomima modelo de regresión polinomial de 4to grado.

a la función de verosimilitud, como sigue

$$\mathcal{L}(\theta|X, \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y} - \Phi(X)^T\theta\|^2}{2\sigma^2}\right) \tag{4.18}$$

Asumiendo la distribución a priori $\pi(\theta) = \mathcal{N}(0, \hat{\Sigma})$, la función a posteriori, vendría determinado, análogamente a (3.24), como

$$\pi(\theta|X, \mathbf{y}) = \mathcal{N}(\overline{\theta}, A^{-1}) \quad \text{donde} \quad \begin{cases} A = \sigma^{-2}\Phi(X)\Phi(X)^T + \widehat{\Sigma}^{-1} \\ \overline{\theta} = \sigma^{-2}A^{-1}\Phi(X)\mathbf{y}. \end{cases} \tag{4.19}$$

Por lo que el mismo análisis planteado por (4.12) y (4.13), la distribución predictiva tomaría la siguiente forma

$$\begin{aligned} f(y^*|\mathbf{x}^*, \mathbf{y}) &= \int f(y^*|x^*, \theta')\pi(\theta'|X, \mathbf{y})d\theta' \\ &= \mathcal{N}\left(\sigma^{-2}\phi(\mathbf{x}^*)^T A^{-1}\Phi(X)\mathbf{y}, \phi(\mathbf{x}^*)^T A^{-1}\phi(\mathbf{x}^*)\right). \end{aligned}$$

por tanto

$$y^*|\mathbf{x}^*, \mathbf{y} \sim \mathcal{N}\left(\sigma^{-2}\phi(\mathbf{x}^*)^T A^{-1}\Phi(X)\mathbf{y}, \phi(\mathbf{x}^*)^T A^{-1}\phi(\mathbf{x}^*)\right). \tag{4.20}$$

Para continuar con los siguientes análisis que vayan acorde con lo que se irá determinando, la expresión para la distribución predictiva, tomará una representación distinta, usando el hecho de que $A = \sigma^{-2}\Phi(X)\Phi(X)^T + \widehat{\Sigma}^{-1}$, por (4.19), y denotando $\tilde{\Sigma} = \Phi(X)^T\widehat{\Sigma}\Phi(X)$.

Veamos, lo siguiente

$$\begin{aligned}
E[y^*|\mathbf{x}^*,\mathbf{y}] &= \sigma^{-2}\phi(\mathbf{x}^*)^T A^{-1}\Phi(X)\mathbf{y} \\
&= \sigma^{-2}\phi(\mathbf{x}^*)^T A^{-1}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T A^{-1}\left[\Phi(X)(\sigma^{-2}\tilde{\Sigma}+\mathbb{I})\right](\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T A^{-1}\left[\Phi(X)\left(\sigma^{-2}\Phi(X)^T\widehat{\Sigma}\Phi(X)+\mathbb{I}\right)\right](\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T A^{-1}\left[\sigma^{-2}\Phi(X)\Phi(X)^T\widehat{\Sigma}\Phi(X)+\Phi(X)\right](\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T A^{-1}\left[\left(\sigma^{-2}\Phi(X)\Phi(X)^T\widehat{\Sigma}+\mathbb{I}\right)\Phi(X)\right](\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T A^{-1}\left[\underbrace{\left(\sigma^{-2}\Phi(X)\Phi(X)^T+\widehat{\Sigma}^{-1}\right)}_{A}\widehat{\Sigma}\Phi(X)\right](\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \\
&= \phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \qquad (4.21)
\end{aligned}$$

$$\begin{aligned}
Var[y^*|\mathbf{x}^*,\mathbf{y}] &= \phi(\mathbf{x}^*)^T A^{-1}\phi(\mathbf{x}^*) \\
&= \phi(\mathbf{x}^*)^T\left[\sigma^{-2}\Phi(X)\Phi(X)^T+\widehat{\Sigma}^{-1}\right]^{-1}\phi(\mathbf{x}^*) \\
&= \phi(\mathbf{x}^*)^T\left[\widehat{\Sigma}+\widehat{\Sigma}\Phi(X)\left(\underbrace{\Phi(X)^T\widehat{\Sigma}\Phi(X)}_{\tilde{\Sigma}}+\sigma^2\mathbb{I}\right)^{-1}\Phi(X)^T\widehat{\Sigma}\right]\phi(\mathbf{x}^*) \\
&= \phi(\mathbf{x}^*)^T\left[\widehat{\Sigma}+\widehat{\Sigma}\Phi(X)\left(\tilde{\Sigma}+\sigma^2\mathbb{I}\right)^{-1}\Phi(X)^T\widehat{\Sigma}\right]\phi(\mathbf{x}^*) \\
&= \phi(\mathbf{x}^*)^T\widehat{\Sigma}\phi(\mathbf{x}^*)+\phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)\left(\tilde{\Sigma}+\sigma^2\mathbb{I}\right)^{-1}\Phi(X)^T\widehat{\Sigma}\phi(\mathbf{x}^*) \\
&\qquad (4.22)
\end{aligned}$$

Por tanto, la distribución de la matriz predictiva es

$$\begin{aligned}
y^*|\mathbf{x}^*,\mathbf{y} \;\sim\; &\mathcal{N}\left(\phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y},\right. \\
&\left.\phi(\mathbf{x}^*)^T\widehat{\Sigma}\phi(\mathbf{x}^*)-\phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\Phi(X)^T\widehat{\Sigma}\phi(\mathbf{x}^*)\right). \qquad (4.23)
\end{aligned}$$

Hay que tener en cuenta que para determinar la varianza (4.22), se ha hecho uso del lema de la inversión matricial (Lema 2.1) para A^{-1}. Además, a la matriz

$$\tilde{\Sigma} = \Phi(X)^T\widehat{\Sigma}\Phi(X) \qquad (4.24)$$

se denomina *matriz Gram*.

4.2.3. Tercer Análisis (Función kernel)

En un caso en particular, supongamos que nuestra función a priori sigue la siguiente ditribución

$$\pi(\theta) \sim \mathcal{N}\left(0, \alpha^{-1}\mathbb{I}\right) \tag{4.25}$$

donde α se conoce como *hiperparámetro* (*no se entrará en detalles sobre este concepto.*). Entonces el análisis de las variables aleatorias Y_i para los valores de salida correspondientes a los $\mathbf{x_i}'s$ (suponiendo los datos de entrenamiento), entonces

$$\begin{aligned} E[h_\theta(\mathbf{x})] &= E\left[\phi(\mathbf{x})^T\theta\right] = \phi(\mathbf{x})^{\mathbf{T}}\mathbf{E}[\theta] = \phi(\mathbf{x})^{\mathbf{T}}\mathbf{0} = \mathbf{0} && (4.26)\\ Cov\left[h_\theta(\mathbf{x_i}), h_\theta(\mathbf{x_j})\right] &= E\left[(h_\theta(\mathbf{x_i}))(h_\theta(\mathbf{x_j}))\right] = E\left[\phi(\mathbf{x_i})^T\theta\theta^T\phi(\mathbf{x_j})\right] \\ &= \phi(\mathbf{x_i})^T\alpha^{-1}\mathbb{I}\phi(\mathbf{x_j}) && (4.27)\end{aligned}$$

Denotando $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \alpha^{-1}\phi(\mathbf{x_i})^T\phi(\mathbf{x_j})$ para todo $i = 1, \ldots, n$, notamos que estos son los elementos de la matriz *matriz Gram* $\tilde{\Sigma}$ (4.24).

En el caso general, con una distribución a priori $\pi(\theta) = \mathcal{N}(0, \widehat{\Sigma})$, definimos la *función kernel* como

$$\begin{aligned} \kappa(\cdot\cdot) : \mathbb{R}^k \times \mathbb{R}^k &\to \mathbb{R} \\ (\mathbf{x}, \mathbf{x}') &\longmapsto \kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\widehat{\Sigma}\phi(\mathbf{x}') \end{aligned} \tag{4.28}$$

la cual evaluada en los n datos de entrenamiento $\mathbf{x}_i$, en datos de prueba (supongamos m) o de forma conjunta, denotamos las siguiente represetanciones matriciales,

$$\tilde{\Sigma} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \ldots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \tag{4.29}$$

$$\Sigma^{**} = \begin{bmatrix} \kappa(\mathbf{x}_1^*, \mathbf{x}_1^*) & \ldots & \kappa(\mathbf{x}_1^*, \mathbf{x}_m^*) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_m^*, \mathbf{x}_1^*) & \ldots & \kappa(\mathbf{x}_m^*, \mathbf{x}_m^*) \end{bmatrix} \quad \Sigma^* = \begin{bmatrix} \kappa(\mathbf{x}_1^*, \mathbf{x}_1) & \ldots & \kappa(\mathbf{x}_1^*, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_m^*, \mathbf{x}_1) & \ldots & \kappa(\mathbf{x}_m^*, \mathbf{x}_m) \end{bmatrix} \tag{4.30}$$

4.3. Regresión por GP's

Los Procesos Gaussianos, comúnmente denominados **GP** o **GP's**, no solo permite el diseño de una técnica de regresión sino que permite complementarse en múltiples técnicas del aprendizaje no supervisado. Un Proceso Gaussiano permite extender distribuciones Gaussianas multivariadas a un espacio de dimensión con la idea de describir funciones sobre distribuciones.

Teniendo el modelo de regresión (4.15), el análisis constructivo desarrollado en el capítulo precedente, un GP'S presenta la siguiente notación

$$h_\theta \triangleq h_\theta(\mathbf{x}) \sim \mathcal{GP}\left(m(x), \kappa(\mathbf{x}, \mathbf{x}')\right) \tag{4.31}$$

donde $m(x)$ y $\kappa(\mathbf{x}, \mathbf{x}')$ se denominan *función media* y *función kernel* correspondientemente, siendo esta última una función de covarianza para dos vectores de características **x** y **x'**. Estas funciones son determinanadas, asumiendo el caso general de una distribución a priori $\pi(\theta) = \mathcal{N}(\mathbf{0}, \widehat{\Sigma})$.

$$\begin{aligned} E[h_\theta] &= \phi(\mathbf{x})E[\theta] = \phi(\mathbf{x})\mathbf{0} = 0 && (4.32)\\ Cov[h_\theta, h'_\theta] &= E[h_\theta h'_\theta] = E\left[\phi(\mathbf{x})^T \theta\theta^T \phi(\mathbf{x}')\right] \\ &= \phi(\mathbf{x})^T \widehat{\Sigma}\phi(\mathbf{x}') = \kappa(\mathbf{x}, \mathbf{x}'). && (4.33) \end{aligned}$$

Por tanto, un proceso gaussianos nos establece una particularidad para $h_\theta\mathbf{x}$ y $h_\theta\mathbf{x}'$ indicando que presentan una distribución conjunta (J-PDF) Gaussiana multivariable con media 0 y covarianza $\kappa(\mathbf{x}, \mathbf{x}')$.

Ahora, entendamos las siguiente notaciones:

$y = h_\theta(\mathbf{x}) + \epsilon$;	
$h_\theta \triangleq h_\theta(\mathbf{x})$;	
$y_i = h_\theta(\mathbf{x}_i) + \epsilon$,	para datos de entrenamiento $(\mathbf{x}_i, y_i)$;
$y^* = h_\theta(\mathbf{x}^*)$,	para un valor de características $\mathbf{x}^*$ puesto a prueba;
$\mathbf{y}$	vector de valores y_i;
$\mathbf{h}_\theta$	vector de valores $h_\theta(\mathbf{x}_i)$ para datos de entrenamiento;
$\mathbf{y}^*$	vector de valores $h_\theta(\mathbf{x}^*) = y^*$ para datos de prueba;
X^*	matriz análoga a (4.12) pero con vectores de prueba $\mathbf{x}^*$

(4.34)

Los procesos gaussianos como se había referido, cada conjunto finito de variables de su conjunto mantiene una distribución multivariable bajo funciones; en (4.32) y (4.33) se determinó la particularidad para dos datos, veamos como se establece esta distribución para una cantidad finita de funciones.

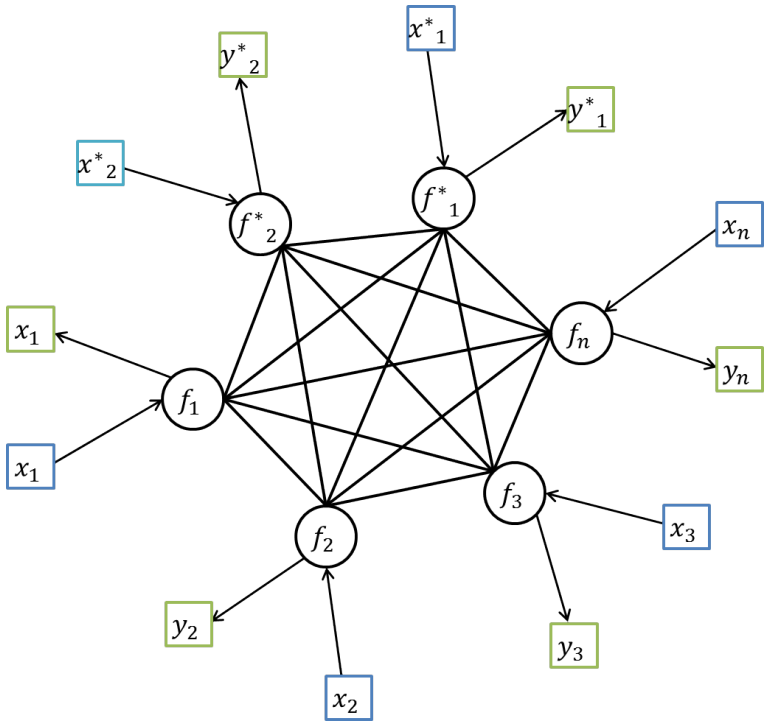


Figura 4.5: Modelo gráfico para el proceso de regresión por procesos Gaussianos GPR. La representación gráfica nos indica conección entre las funciones, que implícitamente dependen de $\mathbf{x}$, donde cada conección establece una distribución conjunta que no depende de las otras funciones, esto nos dice, que la adición de funciones no varía la distribución para cualquier otra distribución conjunta.

$$E[\mathbf{h}_\theta] = E[\Phi(X)^T\theta] = \Phi(X)^T E[\theta] = \Phi(X)^T\mathbf{0} = \mathbf{0}. \tag{4.35}$$

$$\begin{aligned} Cov(\mathbf{h}_\theta) &= E[\Phi(X)^T\theta\theta^T\Phi(X)^T] = \Phi(X)^T E[\theta\theta^T]\Phi(X) \\ &= \Phi(X)^T\widehat{\Sigma}\Phi(X)^T = \tilde{\Sigma}. \end{aligned} \tag{4.36}$$

verificándose que la matriz de covarianza $Cov(\mathbf{h}_\theta) = \tilde{K}$ es la función Gram.
Se logra por tanto una Gaussiana multivariada como distribución conjunta para una cantidad finita de funciones en base a los $\mathbf{x}_i$ para datos de entrenamiento, representada de la siguiente forma

$$\mathbf{h}_\theta \sim \mathcal{N}\left(\mathbf{0}, \tilde{\Sigma}\right). \tag{4.37}$$

Ahora suponiendo un conjunto de funciones en base los vectores de características $\mathbf{x}_i$ para datos de entrenamiento y a $\mathbf{x}^*_i$ para datos de prueba, esto se entiende como el caso general, veamos la representación de su distribución conjunta.

$$
\begin{aligned}
E\begin{bmatrix} \mathbf{h}_\theta \\ y^* \end{bmatrix} &= E\left[\begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\theta\right] = \begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix} E[\theta] = \begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\mathbf{0} = \mathbf{0}. \quad (4.38) \\
Cov\begin{pmatrix} \mathbf{h}_\theta \\ \mathbf{y}^* \end{pmatrix} &= E\left[\begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\theta\left(\begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\theta\right)^T\right] \\
&= E\left[\begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\theta\theta^T\begin{bmatrix} \Phi(X) & \Phi(X^*) \end{bmatrix}\right] \\
&= \begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix} E[\theta\theta^T]\begin{bmatrix} \Phi(X) & \Phi(X^*) \end{bmatrix} \\
&= \begin{bmatrix} \Phi(X)^T \\ \Phi(X^*)^T \end{bmatrix}\widehat{\Sigma}\begin{bmatrix} \Phi(X) & \Phi(X^*) \end{bmatrix} \\
&= \begin{bmatrix} \Phi(X)^T\widehat{\Sigma}\Phi(X) & \Phi(X)^T\widehat{\Sigma}\Phi(X^*) \\ \Phi(X^*)^T\widehat{\Sigma}\Phi(X) & \Phi(X^*)^T\widehat{\Sigma}\Phi(X^*) \end{bmatrix} \\
&= \begin{bmatrix} \tilde{\Sigma} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix} \quad (4.39)
\end{aligned}
$$

denotando $\Sigma^* = \Phi(X^*)^T\widehat{\Sigma}\Phi(X)$ y $\Sigma^{**} = \Phi(X^*)^T\widehat{\Sigma}\Phi(X^*)$.

Se logra obtener, el caso general para una distribución conjunta de funciones, representada como

$$
\begin{bmatrix} \mathbf{h}_\theta \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tilde{\Sigma} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix}\right) \quad (4.40)
$$

En (4.10), se introdujo la distribución predictiva $\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix}$, sin embargo, hasta ahora se ha estructurado (4.40) para enmarcar nuestras ideas.

Además, veamos que haciendo uso de la distribución Gaussiana condiconada, tenemos

$$
y^*|X^*, \mathbf{h}_\theta \sim \mathcal{N}\left(\Sigma^*\tilde{\Sigma}^{-1}\mathbf{h}_\theta, \Sigma^{**} - \Sigma^*\tilde{\Sigma}^{-1}\Sigma^{*T}\right) \quad (4.41)
$$

4.3.1. Distribución Predictiva Conjunta

El modelo de regresión lineal $y = h_\theta(\mathbf{x}) + \epsilon$, representa nuestro estudio para la predicción de nuestros valores de salida, es por ello que extenderemos la visión dada en (4.40) y (4.41), haciendo uso de los valores de salida y_i para los datos de entrenamiento (x_i, y_i), que no se tuvieron en cuenta en pasos anteriores, pues solo se analizó sobre la variable $h_\theta(\mathbf{x}_i)$.

Esto conlleva a relacionar los rudios Gaussianos $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independientes y generados por cada dato de entrenamiento. Además, por la independencia de los ϵ_i, el vector que reúne todos los ruidos Gaussianos ε, determinarían una distribución multivariada, $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I})$.

$$\begin{aligned} E[y] &= E[h_\theta(\mathbf{x}) + \epsilon] = E[\phi(\mathbf{x})\theta + \epsilon] \\ &= \phi(\mathbf{x}) \underbrace{E[\theta]}_{\mathbf{0}} + \underbrace{E[\epsilon]}_{0} = 0. \end{aligned} \tag{4.42}$$

$$\begin{aligned} Co(y, y) &= E\left[(h_\theta(\mathbf{x}) + \epsilon)^2\right] = E\left[h_\theta(\mathbf{x})^2\right] + E\left[\epsilon^2\right] \\ &= E\left[\phi^T(\mathbf{x})\theta\theta^T\phi(\mathbf{x})\right] + \sigma^2 = \phi^T(\mathbf{x})E\left[\theta\theta^T\right]\phi(\mathbf{x}) + \sigma^2 \\ &= \phi^T(\mathbf{x})\widehat{\Sigma}\phi(\mathbf{x}) + \sigma^2 \\ &= \kappa(x, x) + \sigma^2 \end{aligned} \tag{4.43}$$

$$\begin{aligned} Cov(y, y') &= E\left[(h_\theta(\mathbf{x}) + \epsilon)(h_\theta(\mathbf{x}') + \epsilon')\right] = E[h_\theta(\mathbf{x})h_\theta(\mathbf{x}')] + \underbrace{E[\epsilon\epsilon']}_{0} \\ &= E\left[\phi^T(\mathbf{x})\theta\theta^T\phi(\mathbf{x})'\right] = \phi^T(\mathbf{x})E\left[\theta\theta^T\right]\phi(\mathbf{x}') \\ &= \phi^T(\mathbf{x})\widehat{\Sigma}\phi(\mathbf{x}') \\ &= \kappa(\mathbf{x}, \mathbf{x}'). \end{aligned} \tag{4.44}$$

Donde y, y' hemos son independientes, pero podemos generalizar (4.43) y (4.44) usando el *delta de Kronecker* (Definición 2.12), de la siguiente forma

$$Cov(y_a, y_b) = \kappa(\mathbf{x}_a, \mathbf{x}_b) + \sigma^2\delta_{ab}. \tag{4.45}$$

Por tanto, esto nos conlleva a determinar una distribución conjunta sobre funciones, que establezca lo siguiente

$$E[\mathbf{y}] = E\left[\Phi(X)^T\theta + \varepsilon\right] = \Phi(X)^T E\left[\theta\theta^T\right] + E[\varepsilon] = \mathbf{0}. \tag{4.46}$$

$$\begin{aligned} Cov(\mathbf{y}) &= E\left[\mathbf{y}\mathbf{y}^T\right] = E\left[(\Phi(X)^T\theta + \varepsilon)(\Phi(X)^T\theta + \varepsilon)^T\right] \\ &= \Phi(X)^T E\left[\theta\theta^T\right]\Phi(X) + E\left[\varepsilon\varepsilon^T\right] = \Phi(X)^T\widehat{\Sigma}\Phi(X) + \sigma^2\mathbb{I} \\ &= \tilde{\Sigma} + \sigma^2\mathbb{I}. \end{aligned} \tag{4.47}$$

La distribución conjunta se representaría como

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \tilde{\Sigma} + \sigma^2\mathbb{I}\right). \tag{4.48}$$

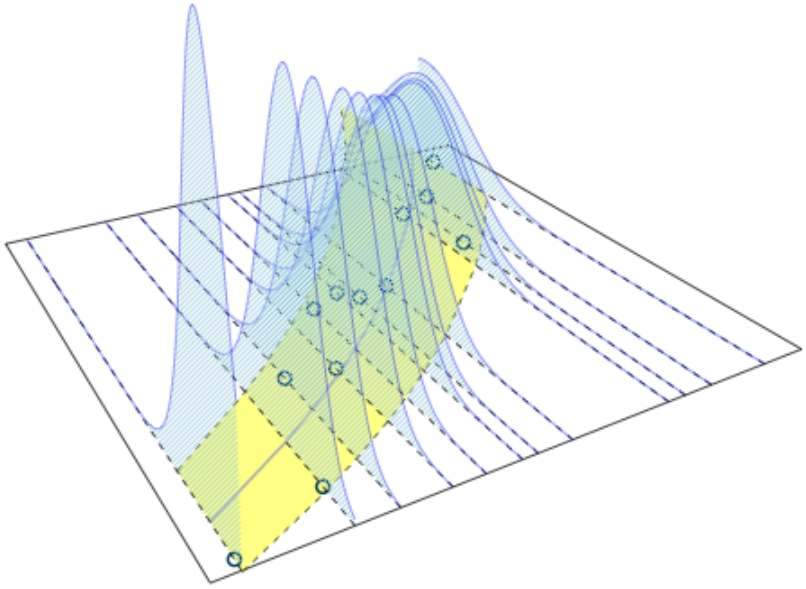


Figura 4.6: Se visualiza la distribución predictiva $y^*|x^*, \mathbf{y}$ para distintos x^*

y usando el mismo proceso para la determinacion de $Cov\begin{pmatrix}\mathbf{h}_\theta \\ \mathbf{y}^*\end{pmatrix}$, tenemos que

$$Cov\begin{pmatrix}\mathbf{y} \\ \mathbf{y}^*\end{pmatrix} = \begin{bmatrix}\tilde{\Sigma}+\sigma^2\mathbb{I} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**}\end{bmatrix} \tag{4.49}$$

Por lo que se establece la distribución correspondiente

$$\begin{bmatrix}\mathbf{y} \\ \mathbf{y}^*\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\tilde{\Sigma}+\sigma^2\mathbb{I} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**}\end{bmatrix}\right). \tag{4.50}$$

Con uso de la distribución Gaussiana condicionada, se obtiene

$$\mathbf{y}^*|X^*, \mathbf{y} \sim \mathcal{N}\left(\Sigma^*(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y}, \Sigma^{**}-\Sigma^*(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\Sigma^{*T}\right). \tag{4.51}$$

la cual se denomina *distribución predictiva conjunta* en este nuevo espacio de funciones, la cual fundamenta la técnica GPR.

4.3.2. Entrenamiento del Modelo (Regresión)

Centrando nuestro objetivo en la determinación de un valor y^* para cierto valor de características $\mathbf{x}^*$, por (4.30), se reduce la distribución (4.50) a una distribución de la forma

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} & \tilde{\Sigma} + \sigma^2\mathbb{I} & & \kappa(\mathbf{x}^*, \mathbf{x}_1) \\ & & & \cdots \\ & & & \kappa(\mathbf{x}^*, \mathbf{x}_n) \\ \kappa(\mathbf{x}^*, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}^*, \mathbf{x_n}) & \kappa(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right). \tag{4.52}$$

Denotando,

$$\kappa^* = \begin{bmatrix} \kappa(\mathbf{x}^*, \mathbf{x}_1) & \cdot & \kappa(\mathbf{x}^*, \mathbf{x}_n) \end{bmatrix} \quad \kappa^{**} = \kappa(\mathbf{x}^*, \mathbf{x}^*) \tag{4.53}$$

Tenemos, haciendo uso de la Proposición 2.4, la distribución predictiva siguiente

$$\begin{aligned} & y^*|\mathbf{x}^*, \mathbf{y} \sim \mathcal{N}\left(\kappa^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y}, \kappa^{**} - \kappa^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}{\kappa^*}^T\right). \\ & \text{donde } \begin{cases} \overline{y}^* = \kappa^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y} \text{ (valor esperado de predicción).} \\ \sigma^2_{y^*} = \kappa^{**} - \kappa^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}{\kappa^*}^T \text{ (varianza de predicción)} \end{cases} \end{aligned} \tag{4.54}$$

Por tanto, se logrará precisar el intervalo de confianza (sección 2.2.8.1) con uso de estos resultados. (ver Figura 4.6)

4.4. Hiperparámetros

En secciones anteriores, los cálculos implicitamente son llevados de acuerdo a como esté definida nuestra función *kernel* (4.28); sin embargo, existe una función sofisticada, usualmente usada en GPR, establecida bajo una base de infinitas funciones (la cual no será estudiada en este proyecto), siendo esta

$$\begin{aligned} \kappa(\cdot\cdot) : \mathbb{R}^k \times \mathbb{R}^k &\rightarrow \mathbb{R} \\ (\mathbf{x}, \mathbf{x}') &\longmapsto \kappa(\mathbf{x}, \mathbf{x}') = \sigma_h^2 \exp\left(-\frac{1}{2l^2}(\mathbf{x} - \mathbf{x}')^2\right) \end{aligned} \tag{4.55}$$

denominada *función exponencial cuadrática*.

Además por (4.48), sabemos que $y \sim \mathcal{N}\left(0, \tilde{\Sigma} + \sigma^2 I\right)$. De todo lo anterior "$l, \sigma_h, \sigma$" se denominan **hiperparámetros**. Por lo que nuestra distribución implicitamente estaría condicionada a dichos parámetros y por tanto

$$f(y|\theta_{hiper}) = \mathcal{N}\left(0, \tilde{\Sigma} + \sigma^2 I\right) = \frac{1}{(2\pi)^{n/2}\left|\tilde{\Sigma} + \sigma^2\mathbb{I}\right|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{y}^T(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y}\right). \tag{4.56}$$

donde $\theta_{hiper} = (l, \sigma_h, \sigma)$.

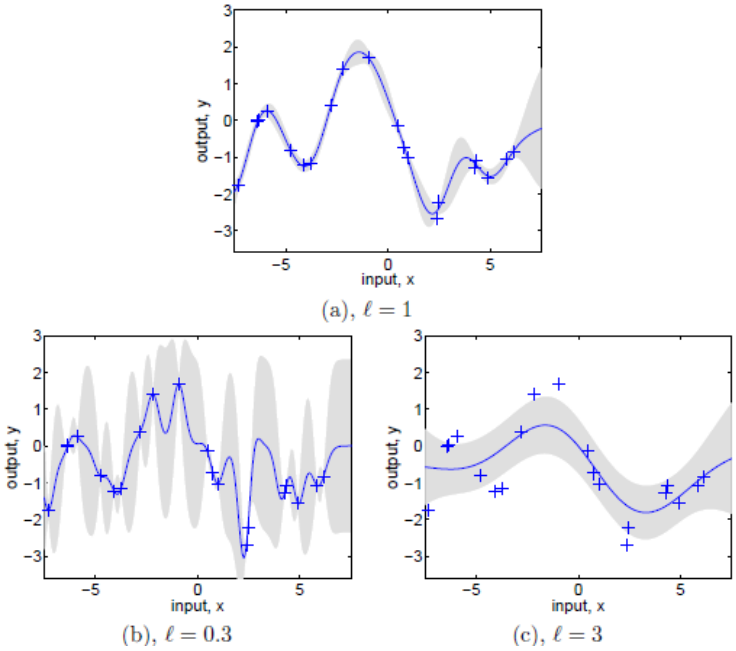


Figura 4.7: Visualización de Regresión por Proceso Gaussianos para ciertos datos usando distintos valores en el hiperparámetro l. He aquí la importancia de la estimación sobre los hiperparámetros.

4.4.1. Maximá verosimilitud

La estadística inferencial nos proporciona el método de *estimación por máxima verosimilitud* (sección 2.2.9.2) para la estimación del parámetro θ_{hiper}.

$$\begin{aligned} \theta_{hiper,\, estimado} &= \max_{\theta_{hiper}} \{l(\theta_{hiper}|\mathbf{y})\} \\ &= \max_{\theta_{hiper}} \left\{ -\frac{1}{2}\ln\left(\left|\tilde{\Sigma}+\sigma^2\mathbb{I}\right|\right) - \frac{n}{2}\ln(2\pi) + \frac{1}{2}\mathbf{y}(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y} \right\} \end{aligned} \quad (4.57)$$

Denotando $B = \tilde{\Sigma} + \sigma^2\mathbb{I}$. Podemos optimizar en base a derivadas parciales, (un resultado del álgebra lineal y optimización) obtiéndose las siguientes ecuaciones

$$0 = \frac{\partial\left(l(\theta_{hiper})\,|\mathbf{y}\right)}{\partial a_i} = \frac{1}{2}Traza\left(B^{-1}\frac{\partial B}{\partial a_i}\right) + \frac{1}{2}\mathbf{y}^T\frac{\partial B}{\partial a_i}B^{-1}\frac{\partial B}{\partial a_i}\mathbf{y}. \quad (4.58)$$

$$(4.59)$$

para $i = 1, 2$ donde $a_1 = \sigma_h, a_2 = l$; y por tanto establecer los hiperparámetros estimados.

4.5. Algoritmo GPR

Algoritmo 2: GPR

Datos: X (matriz definida por los valores de características de n datos de entrenamiento), $\mathbf{y}$ (Vector de valores de salida), σ^2 (varianza de la distribución a priori), κ (función kernel), $\mathbf{x}^*$ (vector de características de un dato de prueba).

1 `/* Matriz Inversa */`

2 $U, S, V \leftarrow SVD\left(\tilde{\Sigma} + \sigma^2\mathbb{I}\right)$

3 $D \leftarrow diag\left(\frac{1}{s_{11}}, \ldots, \frac{1}{s_{nn}}\right)$ `/*` $S = diag(s_{11}, \ldots, s_{nn})$ `*/`

4 $A \leftarrow V^T D U^T$; `/*` $A = inversa\left(\tilde{\Sigma} + \sigma^2\mathbb{I}\right)$ `*/`

5 `/* Valor esperado de predicción. */`

6 $\overline{y}^* \leftarrow \kappa^* A\mathbf{y}$;

7 `/* Varianza de predicción. */`

8 $\sigma^2_{y^*} \leftarrow \kappa^{**} - k^{**} A {\kappa^*}^T$

9 **devolver** $\overline{y}^*, \sigma^2_{y^*}$.

El uso de SVD para determinar la matriz inversa es una de tantas técnicas útiles para mejorar el costo computacional. (ver Anexo pag. 54 - Algortimo SVD)

Capítulo 5

Conclusiones y recomendaciones

En base a como se ha desarrollado este trabajo, se concluye que el análisis bayesiano trae consigo reestructurar la técnica usual de Regresión Lineal en una técnica mucho más sofisticada ofreciéndonos una distribución de datos más acordes con la realidad basando una linealidad sobre funciones e informándonos posteriormente pequeños intervalos con un alto porcentaje de confianza, que podría favorecer ampliamente en programas de simulaciones.

Además, a causa del resultado generado por la técnica de Regresión por Procesos Gaussianos (GPR) podemos entender que si un valor real bajo ciertas características del dato está ciertamente alejado del intervalo de confianza ofrecido por la técnica, se puede intuir la presencia de una anomalía, es decir de un resultado que no va acorde con lo esperado. Es por ello que GPR puede complementarse en un estudio profundo sobre detección de anomalías.

Existen diversas mejoras en técnicas de regresión, sin embargo existen estudios que enfocan los procesos llevados a cabo por la ténica GPR hacia tecnicas de clasificación y de reducción de dimensionalidad, es por ello que el desarrollo de este trabajo no se ve limitado en su objetivo particular, que es el de mejorar la técnica de regresión lineal, sino que es recomendable complementar análisis constructivos que encaminen otro tipo de resultados con objetivo de resolver otro tipo de problemas.

Anexo

Algoritmo 3: SVD

Datos: $A \in \mathbb{R}^{m \times n}$
Resultado: U, S, Vh /* donde $A = USVh$ */
Function *SVD (A):*
  si $n \leq m$ **entonces**
    /* $V = [v_1, \ldots, v_n]$ matriz ortogonal de eigenvectores de $A^T A$, y D lista de eigenvalores λ_i ordenadas de mayor a menor, con correspondencia $v_i \longmapsto \lambda_i$ */
    $V, D \leftarrow eig(A^T A)$;
    /* Sean $\lambda_1, \ldots, \lambda_k$ distintos de cero. */
    $U \leftarrow \left[\frac{Av_1}{s_1}, \ldots, \frac{Av_k}{s_k}\right]$; $S \leftarrow \{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\}$; $Vh \leftarrow [v_1, \ldots, v_k]^T$;
  en otro caso
    /* $V = [v_1 \ldots v_m]$ matriz ortogonal de eigenvectores de AA^T, y D lista de eigenvalores λ_i, con correspondencia $v_1 \longmapsto \lambda_i$ */
    $V, D \leftarrow eig(AA^T)$;
    /* Sean $\lambda_1, \ldots, \lambda_k$ distintos de cero. */
    $U \leftarrow [v_1, \ldots, v_k]$; $S \leftarrow \{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\}$; $Vh \leftarrow \left[\frac{A^T v_1}{s_1}, \ldots, \frac{A^T v_k}{s_k}\right]^T$;
  fin si
  devolver *U, diag(S), Vh* /* diag(S) forma una matriz diagonal */

Bibliografía

[1] Saligrama, V. (2012). *Local Anomaly Detection*. Boston University.

[2] Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology.

[3] Maleki, A. and Do, T (2014). *Review of Probability Theory*. Standford University, United States.

[4] Cinlar, E. (2013). *Introduction to Stochastic processes*. Princeton University, United States.

[5] Anderson, D.; Sweeney, D. y William, T. (2008). *Estadística para Administración y Economía*. University of Cincinnati, United States.

[6] Shanbhag, S. (2007). *Design and Implementation of Parallel Anomaly Detection*. University of Massachusetts, Amherst.

[7] Rasmussen, C.; Williams, C. (2006). *Gaussian Processes in Machine Learning*. Max Planck Institute for Biological Cybernetics, Germany.

[8] Ebden, M. (2008). *Gaussian Processes for Regresion: A Quick Introduction*. Oxford University, United Kingdom.

[9] Belaustegui, C.; Maya, J. (2010). *Distribución Gaussiana Multivariable*. Universidad Javeriana de Bogotá, Colombia.

[10] González, E (2009). *Análisis bayesiano del modelo de regresión lineal con una aplicación a datos astronómicos*. Universidad tecnológica de Mixteca, México.