Técnicas para calcular distribuciones de creencias

Taller Argentino de Computación Científica

Gustavo Landfried

3 de diciembre de 2019

Resumen

Este documento es un complemento de la charla "Técnicas para calcular distribuciones de creencias honestas" y tiene por objetivo servir de herramienta para introducirse en inferencia Bayesiana mediante problemas típicos, de dificultad variable, en sus aspectos prácticos y teóricos.

Elije tu propia aventura

Elija un solo ejercicio (a o b), y resuelvalo.

1. **Flujo de inferencia en modelos generativos** (Sección 1)
   - a) **Implementar** un modelo gráfico utilizando el software `samiam` y determinar cuándo un flujo de inferencia perenece abierto (Subsección 1.2)
     - I Definir los factores del modelo gráfico visto en clase
     - II Extender el modelo gráfico
     - III Observar el efecto de los observables sobre las creencias a priori y a posteriori
     - IV Determinar cuándo un flujo de inferencia permanece abierto
   - b) **Derivar** algunas de las creencias a priori y creencias a posteriori mediante las reglas de la probabilidad (Subsección 1.3)
2. **Habilidad en la industria del video juego** (Sección 2)
   - a) **Implementar** una estimación de habilidad sobre jugadores en el lenguaje de programción de su preferencia (Subsección 2.2)
     - I Utilizar la librería TrueSkill de python o la respectiva a su lenguaje de programación
     - II Simular un jugador con oponentes al azar utilizando el modelo generativo
     - III Estimar la habilidad del jugador, considerando conocidos a los opnentes, $\sigma = 1$
     - IV Repetir el punto ii y iii, y observar como varían las observaciones
   - b) **Derivar** la verosimilitud, la evidencia y la posterior del modelo TrueSkill usando el sum-product algorithm sobre su factor graph (Subsección 2.3)
3. **Regresión lineal Bayesiana** (Sección 3)
   - a) **Implementar** una selección de modelo de regresión lineal Bayesianos sobre datos simulados en el lenguaje de programación de su preferencia (Subsección 3.2)
     - I Definir la posterior, la verosimilitud y la evidencia,
     - II Simular datos con ruido provenientes de una sinoidal
     - III Ajustar regresiones polinomiales de grado 0 hasta 9 a los datos
     - IV Seleccionar modelo basado en la evidencia
   - b) **Derivar** la distribución de creencias a poteriori y la evidencia de la regresión lineal Bayesiana (Subsección 3.3)

1. Flujo de inferencia en modelos generativos

1.1. Introducción

Hemos visto que las reglas óptimas de razonamiento en contexto de incertidumbre son las llamadas "reglas de la probabilidad", la regla de la suma y regla del producto. Estas son las únicas reglas que necesitamos para actualizar nuestras distribuciones de creencias. Pero para eso necesitamos definir nuestro modelo generativo. Si planteamos un modelo "causal" entre las variables para los que tenemos definidas sus distribuciones de creencia a priori, estaremos proponiendo lo que se conoce como modelo generativo. El enfoque generativo es el más completo pues permite definir la distirbución de probabilidad conjunta y condicional, generar datos de esa distribucón de probabilidad y separar la etapa de inferencia de la etapa de decisión.

Variables del modelo "entrevista de trabajo":

a) La **admisión** al trabajo depdende del resultado r de la entrevista

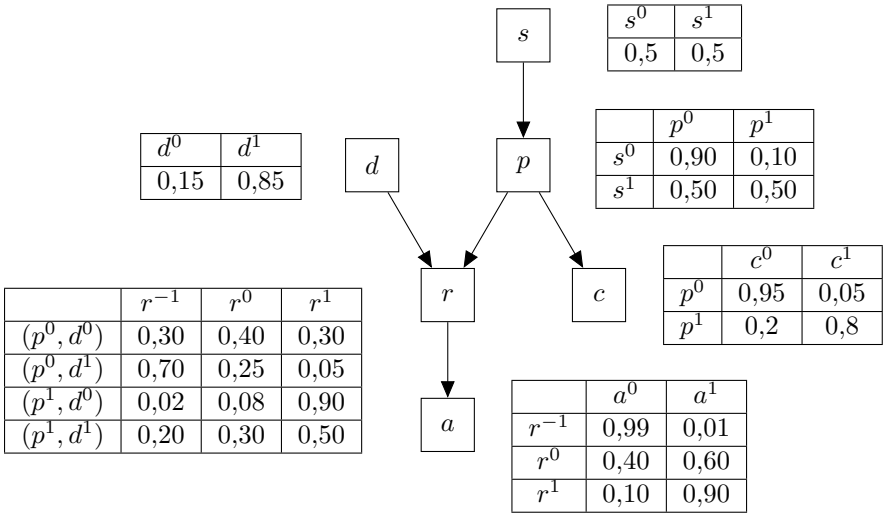
r) El **resultado** depende de la dificultad d de la entrevista y el rendimiento p de la peronsa

d) La **dificultad** es independiente. La mayoría de las veces es difícil, a veces fácil

p) El **rendimiento** depende del nivel socio-económico s

s) El **nivel socio-económico** de quienes se presentan está dividido en mitades iguales

c) **Información externa** que depende del rendimiento, como puede ser la estimación Elo de habilidad en una página de juegos en línea



s^0	s^1
0,5	0,5

d^0	d^1
0,15	0,85

	p^0	p^1
s^0	0,90	0,10
s^1	0,50	0,50

	r^{-1}	r^0	r^1
(p^0, d^0)	0,30	0,40	0,30
(p^0, d^1)	0,70	0,25	0,05
(p^1, d^0)	0,02	0,08	0,90
(p^1, d^1)	0,20	0,30	0,50

	c^0	c^1
p^0	0,95	0,05
p^1	0,2	0,8

	a^0	a^1
r^{-1}	0,99	0,01
r^0	0,40	0,60
r^1	0,10	0,90

Al definir un modelo gráfico estamos definiendo una distribución de probabilidad $P(s, d, p, r, c, a)$.

1.2. Flujos de inferencia

¿Qué pasa con nuestras creencia sobre las diferentes variables ocultas cuando observamos algunas de ellas?. Si bien podríamos resolver este problema a mano, aplciando las reglas de la suma y del producto, en este primer ejercicio el objetivo es implementar el modelo gráfico en algún software que nos permita jugar con los distintos flujos de inferencia y verificar rápidamente si la veracidad de la siguiente tabla. El objetivo es adquirir intución respecto de la apertura de los flujos de inferencia en los modelos causales.

	$V \notin$ Observable	$V \in$ Observable
$X \to V \to Y$	Sí	No
$X \leftarrow V \leftarrow Y$	Sí	No
$X \leftarrow V \to Y$	Sí	No
$X \to V \leftarrow Y$	$\underbrace{\text{No}}_{\text{Y ningún descendiente observable}}$	$\underbrace{\text{Sí}}_{\text{O algún descendiente observable}}$

Pueden utilizar el software de su preferencia. Sin embargo, para esta tarea proponemos el software `samiam` dado que es muy útil para visualizar el efecto de la evidencia sobre los distribuiciones de creencia. Lo pueden descargar en `http://reasoning.cs.ucla.edu/samiam/` y no requiere permisos especiales.

Si deciden usar el software `samiam`, en el repositorio hay un archivo con el modelo implementado parcialmente. Solamente falta agregar la variable "nivel socio-económico" (s), la relación causal sobre el rendimiento (p) y actualizar las distribución condicionales de ambas variables.

Una vez editado el modelo, pasen a modo "inferencia" seleccionando `Mode>Query Mode`. Abran todas las distribuciones de creencias seleccionando `Query>Show monitors>Show All` y diviertanse viendo eligiendo observables.

1.3. Reglas de la suma y el producto

Para entender el motivo por el cual los flujos se abren y se cierran es un buen ejercicio calcular las creencias a priori y a posteriri aplicando a mano las reglas de la suma y el producto. La regla de la suma dice que nuestra creencia marginal se puede calcular integrando la distribución de creencias conjunta.

$$P(X) = \sum_{Y} P(X, Y) \tag{1}$$

La regla del producto dice que la distribución de creencia conjunta se puede expresar como una multiplicación de distribuciones creencia unidimensionales.

$$P(X, Y) = P(Y|X)P(X) \tag{2}$$

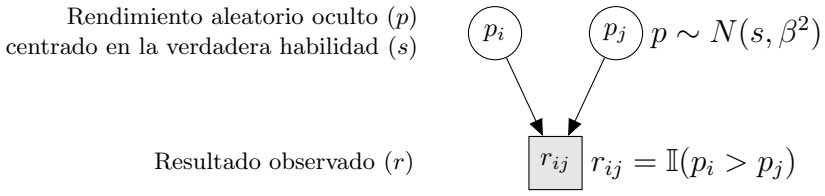
Elija usted las creencia a priori y a posterior que quiera. Si no se le ocurre ninguna, le proponemos computar lo siguiente:

- $P(d^0|c^0)$
- $P(d^0|c^0, a^1)$
- $P(a^1|p^0)$
- $P(a^1|c^0, p^0)$

2. Habilidad en la industria del video juego

2.1. Introducción

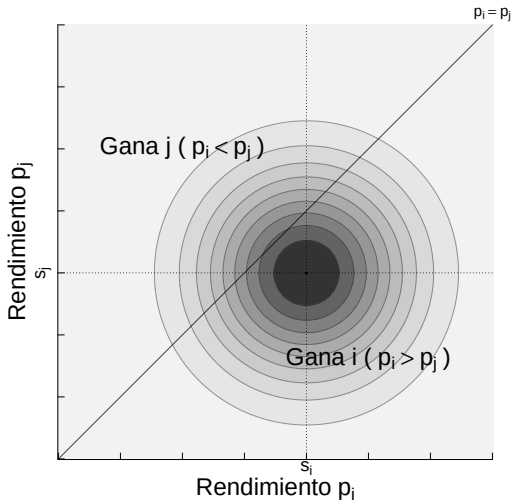
TrueSkill es el estado del arte para estimación de habilidad en la industria del video juego. Es una versión Bayesiana del modelo desarrollado por Arpad Elo en 1959, usado por la federación internacional de ajdrez desde 1970. La idea principal consiste en considerar los resultados observables r en función de rendimientos ocultos p de los jugadores, variables aleatorias centradas en la verdadera habilidad s.



El modelo supone que la persona ganadora es aquella que tuvo mayor rendimiento en la partida. Esto permite inferir quién tuvo mayor rendimiento y por lo tanto calcular la probabilidad de que ese evento ocurra. En términos gráficos, la probabilidad del resultado dada las estimaciones previas es

$$P(p_i > p_j | s_i^{\text{old}}, s_j^{\text{old}}) \tag{3}$$

La probabilidad de ganar de un jugador y de otro, en términos gráficos, es equivalente al volumen debajo de la curva de una lado y del otro del recta $p_i = p_j$



2.1.1. Estimación puntual

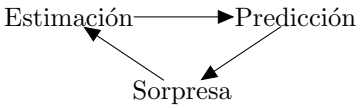
La solución propuesta por Elo fue comenzar con una estimaciones inciales arbitrarias, que son luego son actualizadas de la siguiente forma,

$$s_i^{\text{new}} = s_i^{\text{old}} + K\Delta$$

donde,

$$\Delta = \underbrace{(2\,\mathbb{I}(p_i > p_j) - 1)}_{\substack{\textbf{Signo} \\ \text{del resultado}}} \underbrace{(1 - P(p_i > p_j | s_i, s_j))}_{\substack{\textbf{Sorpresa} \\ \text{del resultado}}}$$

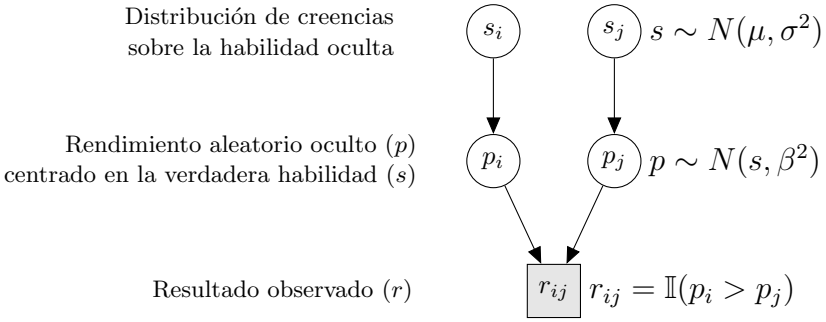
El modelo de solución del sistema Elo



El problema principal es la falta de incertidumbre respecto de las estimaciones. La estimación inicial arbitraria no puede tener el mismo estatus que aquella estimada luego de varias partidas. Este problema fue resuleto mediante la constante K. Sin ella, lo que pierde un jugador lo gana el otro. En la práctica a jugadores con muchas partidas se le asigna K bajos, y a jugadores con pocas partidas valores altos. Sin embargo, esta es una solución ad-hoc.

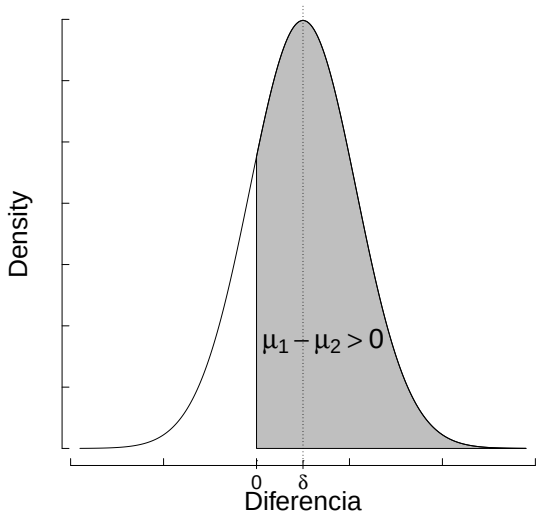
2.1.2. Distribución de creencias

Un tratamiento correcto de la incertidumbre necesariamente requiere la incopración de una distribución de creencias sobre las habilidades.



Una distribución de creencias a priori con poca información deberá tener mucha varianza alrededor de la media de habilidades de la población. Del mismo modo que en el modelo gráfico discreto visto en la sección anterior, las distribuciones de creencia a posteriori y la evidencia no son más que las consecuancia de aplicar las reglas de la suma y el producto. Usando las reglas de la probabilidad podemos probar que la evidencia, o predicción a priori del dato observado, es

$$P(r_{12} = 1|s_1, s_2) = 1 - \Phi(0| \underbrace{\mu_1 - \mu_2}_{\substack{\text{Diferencia} \\ \text{esperada}}\ \delta}, \underbrace{\sigma_1^2 + \sigma_2^2 + 2\beta^2}_{\substack{\text{Varianza} \\ \text{total}}\ \vartheta^2}) \tag{4}$$



y en caso perdedor

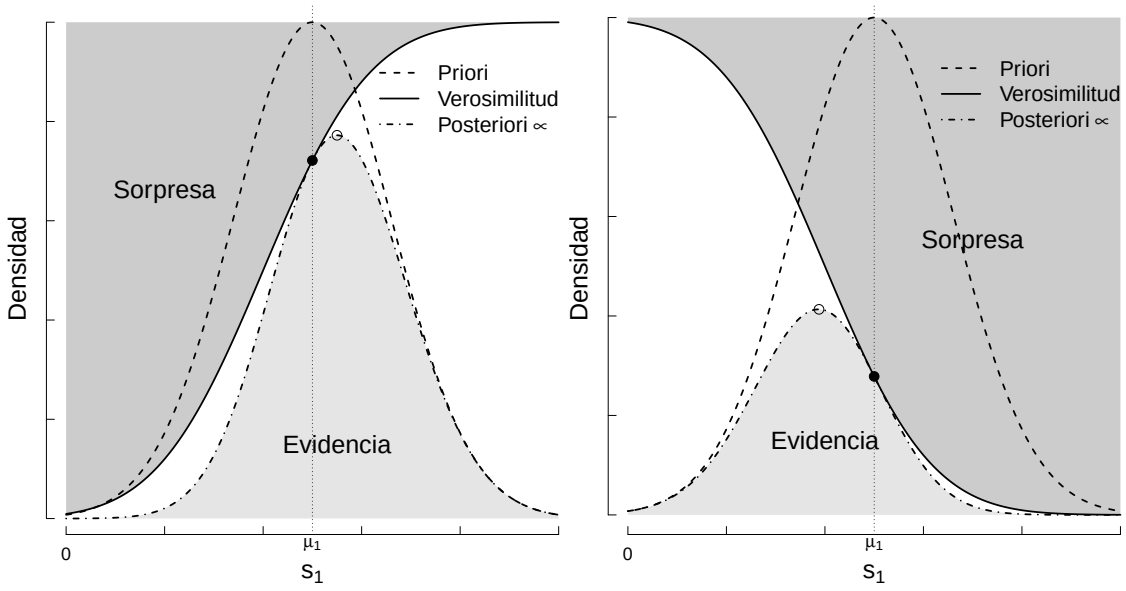
$$P(r_{12} = 0|s_1, s_2) = \Phi(0|\delta, \vartheta^2) \tag{5}$$

Luego de ver un resultado, la posterior analítica para el caso canador es

$$\overbrace{P(s_1 \mid r)}^{\text{Posteriori}} \propto \overbrace{N(s_1 \mid \mu_1, \sigma_1^2)}^{\text{Priori}} \overbrace{1 - \Phi(0 \mid s_1 - \mu_2, \vartheta^2 - \sigma_1^2)}^{\text{Verosimilitud}} \quad \text{Caso ganador} \tag{6}$$

donde *Phi* representa la acumulada de la distribución Gaussiana. Del mismo modo, la evidencia para el caso perdedor es,

$$\overbrace{P(s_2 \mid r)}^{\text{Posteriori}} \propto \overbrace{N(s_2 \mid \mu_2, \sigma_2^2)}^{\text{Priori}} \overbrace{\Phi(0 \mid \mu_1 - s_2, \vartheta^2 - \sigma_1^2)}^{\text{Verosimilitud}} \quad \text{Caso perdedor}$$



(a) Caso ganador (b) Caso perdedor

Notar que la distirbución a priori, por ser una distribución de probabilidades, integra 1. Y que la distribución proporcional a la posteriori integra la evidencia. Además notar que la verosimilitud, al ser la acumulada de la distribución Gaussiana, va de 0 a 1 y cumple la función de filtro de la distribución de creencias a priori. En las regiones donde la verosimilitud es 1, la posterior y el prior permanecen iguales. En las regiones donde la verosimilitud es 0, la posterior se anula. Si bien a simple vista la distribución a posteriori aparenta se una distribución Gaussiana, por su asimetría sabemos que no lo es.

Aproximación variacional de la posterior Si bien podemos calcular la distribución a posteriori exacta, en la práctica es preferible aproximarla por una distribución Gaussiana de modo de tener una solución analítica eficiente. A este perspectiva de aproximar la posterior mediante una distribuciones que pertenece a cierta familia, se la conoce como variacional.[1]

2.2. Estimación de habilidades sintéticas

La mayor parte de los lenguajes de programación tienen alguna librería que implementa la solución variacional del modelo de habilidad, generalmente denominada `trueskill`. Para ganar conocer el comportamiento de TrueSkill proponemos realizar un experimento con datos simulados. El objetivo es verificar cuánto tarda TrueSkill en estimar la verdadera habilidad y cuál es la variabilidad de sus estimaciones.

Para ello proponemos el siguiente diseño experimental:

μ_1) Definimos al comienzo la habilidad real del jugador forcal, algún valor entre $\mu_1 \in \{15, 20, 30, 35\}$

μ_2) El cual se va a enfrentar a 50 oponentes al azar con habilidad real ($\mu_2 \sim N(25, \frac{25}{3})$) en partidas 1 vs 1

[1] La aproximación variaicional es uno de los 3 métodos para computar distribuciones de creencia a posteriori, junto a las soluciones analíticas exactas y los métodos de sampleo. Debido a su versatilidad y eficiencia es el método utilizado para desarrollar redes neuronales Bayesianas. Ver los videos Summer School on Deep Learning and Bayesian Methods organizados por Deep Bayes (link).

β, p) Cuando compiten, ambos "tiran la moneda de su rendimiento" de una distribución Gaussian centrada en su propia habilidad y con cierto rudio, $p \sim N(\mu, \beta)$. Probar $\beta = \frac{25}{6}$.

r_{12}) El resultado (ganador o perdedor) que surge de determinar quién obtuvo mayor rendimiento, se utilizará para actualizar la distribución de creencias sobre las habilidades

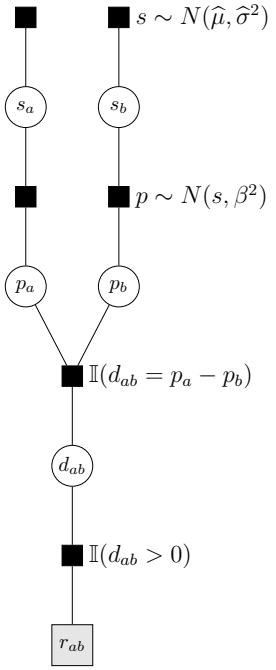
σ_1) La incertidumbre inicial sobre la habilidad del jugador focal queremos que sea alta, por ejemplo $\sigma_1 = \frac{25}{3}$

σ_2) La incertidumbre respecto de la habilidad de los oponentes la vamos a considerar baja, por ejemplo $\sigma_2 = 1$

Para ver la velocidad de convergencia de y la variabilidad en la estimación, repetir el experimento 50 veces.

2.3. Sum-product algorithm

La actualización de creencias relactivas al modelo de habilidad Bayesiano depende de las reglas de la suma y el producto. El sum-product algorithm es una técnica eficiente de pasaje de mensajes para computar distribuciones de creencias a posteriori a partir factor graph, i.e. modelos gráficos en el que las probabilidades condicionales aparecen como un nodo diferenciado al nodo variable típco de las redes bayesianas.



Las marginales en un factor graph se calculan como el producto de los mensajes recibidos por los nodos vecinos

$$P(x) = \prod_{h \in n(x)} m_{h \to x}$$

donde

$m_{x \to f}(x)$: Mensaje de variable x a factor f

$m_{f \to x}(x)$: Mensaje factor f a variable x

$n(v)$: Conjunto de nodos vecinos del nodo v

$$m_{x \to f}(x) = \prod_{h \in n(x) \setminus \{f\}} m_{h \to x}(x)$$

$$m_{f \to x}(x) = \int \Big(f(X) \prod_{h \in n(f) \setminus \{x\}} m_{h \to f}(h) \Big) dX_{\setminus \{x\}}$$

Para calcular la evidencia y la posterior del modelo de habilidad es suficiente con aplicar el pasaje de mensajes definido por el sum-product algorithm junto a las siguientes propiedades

Simetría: $N(x|\mu, \sigma^2) = N(\mu|x, \sigma^2) = N(-\mu| - x, \sigma^2) = N(-x| - \mu, \sigma^2)$

Estandar: $N(x|\mu, \sigma^2) = N(\frac{X-\mu}{\sigma}|0, 1)$

Acumulada: $\frac{\partial}{\partial x} \Phi(x|\mu, \sigma^2) = N(x|\mu, \sigma^2)$

Indicadora: $\iint_{-\infty}^{\infty} \mathbb{I}(x = h(y, z))\, f(x)\, g(y)\, dx\, dy = \int_{-\infty}^{\infty} f(h(y, z))\, g(y)\, dy$

Producto: $\int_{-\infty}^{\infty} N(x|\mu_x, \sigma_x^2) N(x|\mu_y, \sigma_y^2)\, dx = \int_{-\infty}^{\infty} \underbrace{N(\mu_x|\mu_y, \sigma_x^2 + \sigma_y^2)}_{\text{constante}} \underbrace{N(x|\mu_*, \sigma_*^2) dx}_{\text{integra 1}}$

3. Modelo lineal Bayesiano

3.1. Introducción

Cuando hablamos de modelos lineal nos referimos a funciones $\boldsymbol{t} = f(\boldsymbol{w}, \boldsymbol{x}) + \epsilon$ que son lineales en sus parámetros $\boldsymbol{w}$, no en sus observables $\boldsymbol{x}$. Utilizando transformaciones no lineales sobre los observables, $\boldsymbol{\phi}(\boldsymbol{x})$, es posible modelar cualquier tipo de relación no lineal entre los observables $\boldsymbol{x}$ y los no observables $\boldsymbol{t}$ mediante funciones lineales en los parámetros $\boldsymbol{w}$.

El modelo lineal es un modelo generativo. Propone la existencia de una relación causal $y(\cdot)$ entre una variable de interés (target) t y los vectores de variables observables $\boldsymbol{x}$ y no observables $\boldsymbol{w}$. En terminos generales la relación causal, $y(\cdot)$ se define como,

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=0}^{M-1} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{7}$$

Donde el vector $\boldsymbol{\phi}(\mathbf{x}_1) = (\phi_0(\boldsymbol{x}_1), \phi_1(\boldsymbol{x}_1), \ldots, \phi_{M-1}(\boldsymbol{x}_1))^T$ son la M tranformaciones no-lineales. Notar que cada función de base $\phi_j(\cdot)$ recibe el vector-input completo $\boldsymbol{x}_i$. Hoy trabajaremos con una única dimensión, por lo que el vector $\boldsymbol{x}_i$ será simplemente un escalar. Aquí usamos la convención $\phi_0(\boldsymbol{x}) = 1$. El modelos lineal más simple es aquel que también es lineal en sus variables observables $\boldsymbol{x}$. En este caso la transformaciones no es más que la identidad $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$.

Además de la parte estrictamente causal, el modelo lineal considera siempre la existencia de un factor aleaotio ϵ. El modelo completo que relaciona la variable de interés t con las variables observables $\boldsymbol{x}$ se compone de ambos términos.

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{8}$$

El ruido aleatorio ϵ lo vamos a modelar proviniendo de una distribución Gaussiana, centrada en cero y precisión (inversa de la varianza) β, $\epsilon \sim N(0, \beta^{-1})$. [2]

Dado que tenemos una sola componente determinista $y(\cdot)$ y una sola componenete aleaotio ϵ, el modelo generativo de las variables de interés t es,

$$P(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) = \mathcal{N}(t|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \beta^{-1}) \tag{9}$$

Generalmente se supone que las variables de interés t son independiente e identicamente distribuida, por lo que la probabilidad conjunta de un vector de variables de interés $\boldsymbol{t}$ se obtiene como la multiplicación de cada uno de los términos individuales

$$P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \tag{10}$$

También podemos representar N distirbuciones Gaussianas independientes a través de una única distirbución Gaussiana multivariada, en la que la matriz de covarianzas con valores no nulos en la diagonal.

$$P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{w}^T \boldsymbol{\Phi}, \beta^{-1}\mathbf{I}) \tag{11}$$

Donde $\mathbf{I}$ es la matriz identidad y $\boldsymbol{\Phi}$ se la matriz de diseño

[2] Una pregunta que surge naturalmente es plantearse si las decisiones que tomamos para modelar la relaciones entre variables es la correcta.En este punto no hay que perder de vista que los modelos no son más que representaciones de la realidad, así como los mapas no son el territorio. Esto no significa que sean todas igualmente falsas y desechables. Hay representaciones mejores que otras. Y la inferencia Bayesiana proveé una forma de computar la creencias óptimas sobre los modelos dada la evidencia. Entonces terminemos de desarrollar la estrucutra de la familia de modelos lineales y dejemos la selección de modelos para después.

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^T \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^T \end{pmatrix} \tag{12}$$

3.1.1. Estimación clásica

La solución clásica, o frecuentista, elige los parámtro $\boldsymbol{w}$ que tienen máxima verosimilitud.

$$\boldsymbol{w}_{\text{MV}} = \max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) \tag{13}$$

Los prámetros $\boldsymbol{w}$ que que minimizan la distancia entre la curva $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i)$ y los datos $\boldsymbol{t}$ son los que tienen mayor verosimilitud (demostración en el anexo).

$$\max_{\boldsymbol{w}} P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \min_{\boldsymbol{w}} \sum_{i=1}^{n} (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \tag{14}$$

Aumentar la flexibilidad de la curvas $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i)$ a través de modelos más complejos, e.g. el grado de los modelos polinomiales, siempre permite reducir la distancia a los datos. Este enfoque conduce a un grave problema conocido como over-fitting (o sobreajuste). Las estrategias comunes para evitar el sobreajuste es agregar términos de regularización ad-hoc a la función de error o través de la evaluación de la función de costo en conjuntos de datos de validación.

3.1.2. Estimación Bayesiana

Aquí veremos como un tratamiento Bayesiano de la regresión linea no solo evita el over-fitting que surge del criterio de estimación puntual basado en máxima verosimilitud, sino que ofrece una forma natural para determinar la complejidad del modelo usando tan solo los datos de entrenamiento.

Antes de computar nuestra distribución de creencias a posteriori sobre los parámetros $\boldsymbol{w}$ es necesario antes definir nuestra distribución de creencias a priori.

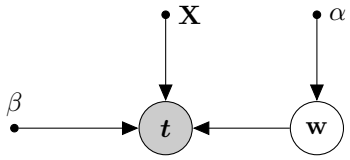
$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{15}$$

Dada la elección de una distribución Gaussiana a priori, la distribución a posteriori también será Gaussiana. Podemos derivar este resultado mediante el procedimiento usal de completar los cuadrados en el exponente, y encontrando el coefficient de normalización. El resultado general de este procedimiento se encuentra en la ecuación (22) y es lo que se necesita para derivar la posterior y la evidencia. La distribución posteriori sobre $\mathbf{w}$ es

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^T\mathbf{t} \tag{16}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{17}$$

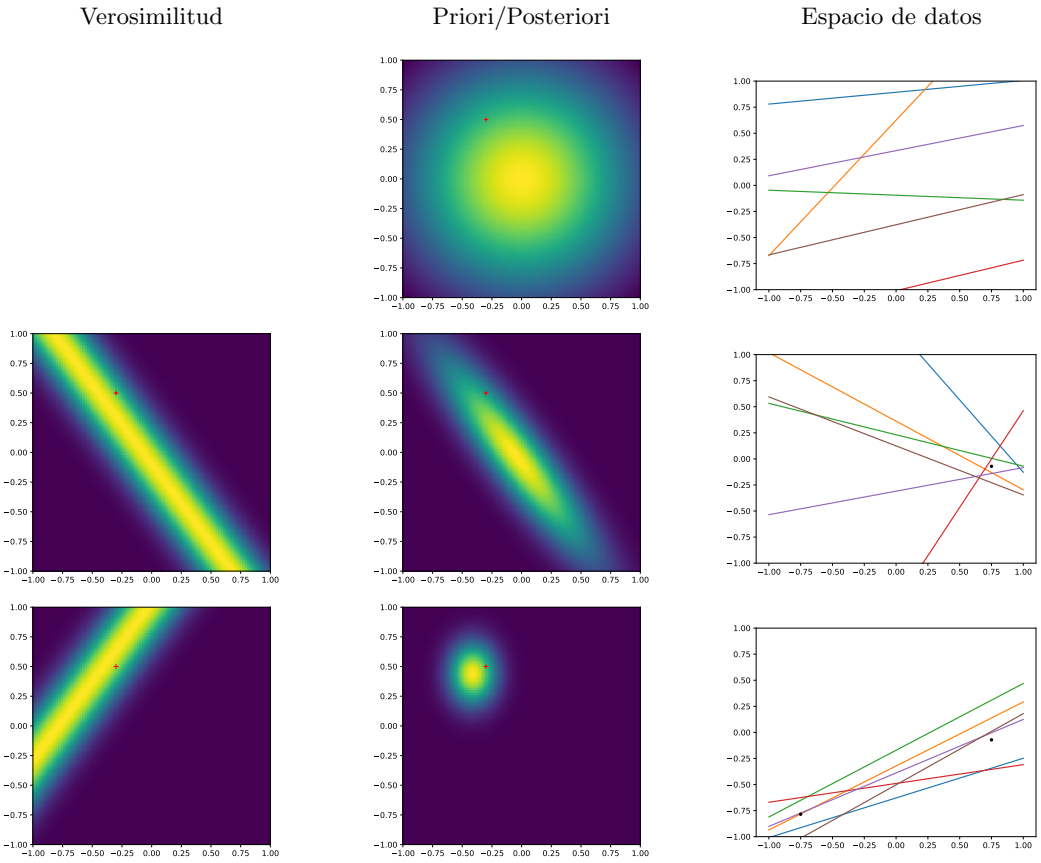
Para tener una solución analítica vamos a tratar a β como constante conocida. Además, en los problemas de regresión no buscamos inferir la distribución del vector de entrada $\mathbf{x}$, por lo que lo podemos tratar como constantes conocidas. En términos gráficos el modelo es,



Y la evidencia del modelo es

$$P(t) = \mathcal{N}(\boldsymbol{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) \tag{18}$$

A modo de visualización, mostramos la actualización de creencias a medida que van llegando datos de una función objetivo lineal ajustada utilizando la identidad como tranformación de base.



3.2. Selección de modelo

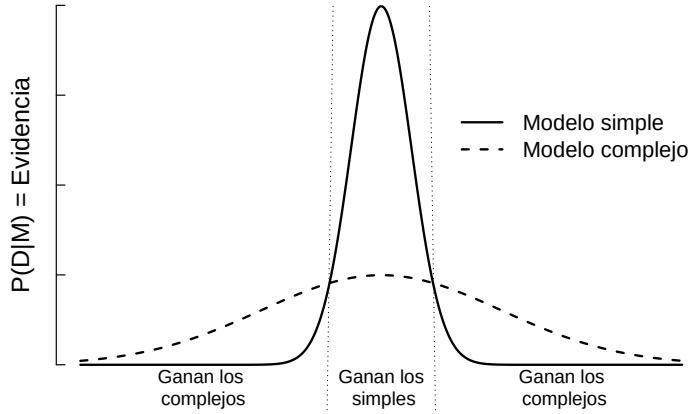
En el ejemplo anterior ajustamos una función lineal con un modelo lineal. En general no conocemos la función objetivo. Y para ajustarla tenemos muchos modelos alternativos que podríamos usar. Para decidir sobre los distintos modelos, lo mejor que podemos hacer es plantear nuestra distribución de creencias sobre los modelos luego de haber visto los datos.

$$P(\mathrm{M}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{M})P(\mathrm{M})}{P(\mathrm{D})}$$

En este caso no es posible calcular el denominador $P(\mathrm{D})$. Sin embargo, sí podemos comparar modelos.

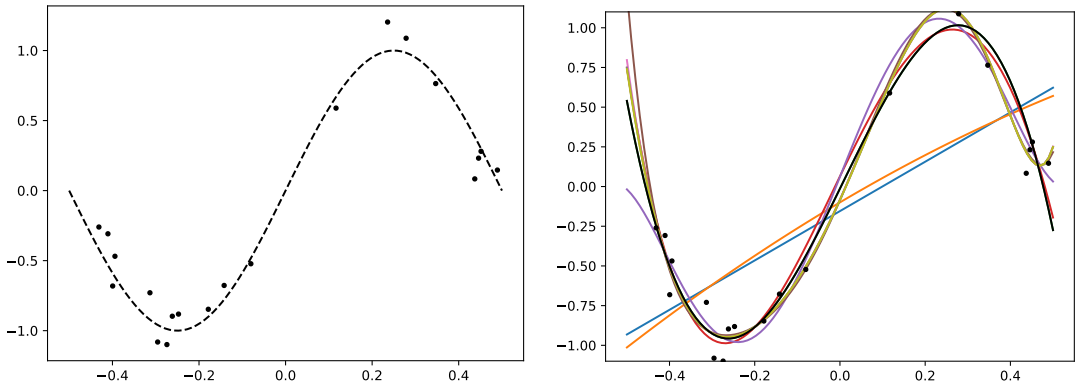
$$\frac{P(\mathrm{M}_i|\mathrm{D})}{P(\mathrm{M}_j|\mathrm{D})} = \underbrace{\frac{P(\mathrm{D}|\mathrm{M}_i)}{P(\mathrm{D}|\mathrm{M}_j)}}_{\text{Evidencia!}} \frac{P(\mathrm{M}_i)}{P(\mathrm{M}_j)}$$

En caso de no tener preferencia a priori sobre ningún modelo, podemos comparar modelos calculando la predicción conjunta de los datos observados, es decir, de la evidencia. La evidencia, al ser una distribución de probabilidad que siempre integrar 1, sufre una penalización natural a medida que se le agregan dimensiones a los modelos.



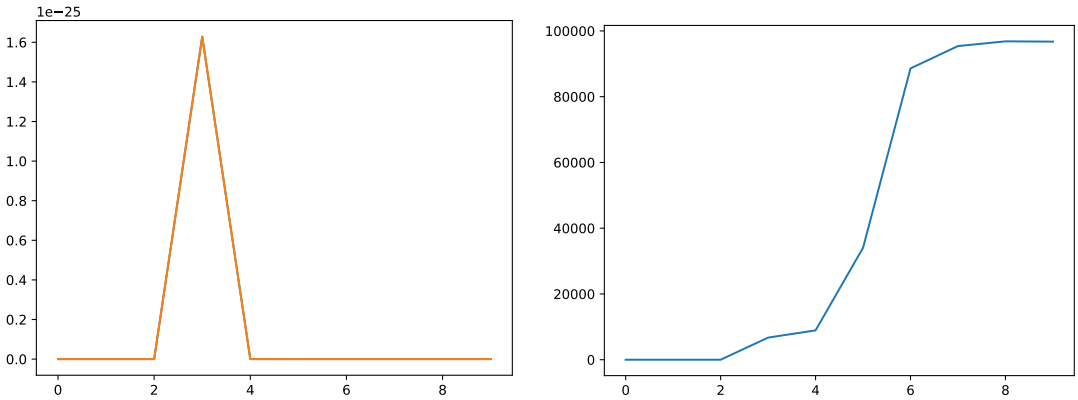
Para verificar la útilidad de la evidencia (predicción a priori de los datos observados), realizar el siguiente experimento.

- Generar datos de una sinoidal en el intervalo con poco ruido, por ejmplo con precisión $\beta = 25$
- Ajustar regresiones polinomiales de grado 0 hasta 9 con mucha incertidumbre a priori, por ejemplo precisión $\alpha = 10^-7$



Y luego

- Seleccionar modelo basado en la evidencia
- Comparar el comportamiento de la evidencia respecto al de máxima verosimilitud



3.3. La base de muchos algoritmos

La regresión lineal basada en tranformaciones no lineales se encuentra en la base de los algoritmos más importantes del área de aprendizaje automático e inteligencia artificial, lo cuales pueden clasificarse del siguiente modo:

1. basadas en transformaciones no lineales fijas
2. basadas en transformaciones no lineales adaptativas
3. basadas en una jerarquía de transformaciones no lineales adaptativas

Las redes neuronales profundas son un ejemplo de regresiones lineales basadas en una jerarquía de transformaciones adapativas. Comprender a cabalidad la regresión lineal básica, aquella basada en transformaciones fijas, es de gran ayuda para comprender aspectos que aprecen en todos los algoritmos del área de aprendizaje automático e inteligencia artificial.

Una propiedad importante del modelo lineal basado en distribución Gaussiana multivariada es la siguiente.

Dado una distirbución Gaussiana marginal $p(\mathbf{x})$ y una distribución Gaussiana condicional en donde $P(\mathbf{y}|\mathbf{x})$ tiene como media una función lineal sobre $\mathbf{x}$.

$$P(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{19}$$

$$P(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{20}$$

Luego,

$$P(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{21}$$

$$P(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}[\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}],\, \boldsymbol{\Sigma}) \tag{22}$$

donde,

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{23}$$

Si quieren entender como se llega a esta, recomiendo leer el capítulo 2 del libro de Bishop: "Pattern Recognition and Machine Learning".

Utilicen este resultado para derivar la evidencia y la posterior del modelo lineal Bayesiano.

4. Anexo

4.1. Máxima verosimilitud

$$\begin{aligned}
\log p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w},\beta) &= \sum_{i=1}^{n} \log\, N(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i),\sigma) \\
&= \sum_{i=1}^{n} \log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} = \sum_{i=1}^{n} \log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \log\, e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} \\
&= n\log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \log\, e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} = n\log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}} \\
&= n\, \log\, \sqrt{\beta} - n\, \log\, \sqrt{2\pi} - \frac{\beta}{2}\sum_{i=1}^{n}(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \\
&\propto -\sum_{i=1}^{n}(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2
\end{aligned} \tag{24}$$

4.2. Multiplicación de normales

Luego, el problema que tenemos que resolver es

$$\int N(x;\mu_1,\sigma_1^2)N(x;\mu_2,\sigma_2^2)dx \tag{25}$$

Por defnición,

$$\begin{aligned}
N(x;y,\beta^2)N(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\underbrace{\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}_{\theta}\right)
\end{aligned} \tag{26}$$

Luego,

$$\theta = \frac{\sigma_2^2(x^2+\mu_1^2-2x\mu_1)+\sigma_1^2(x^2+\mu_2^2-2x\mu_2)}{2\sigma_1^2\sigma_2^2} \tag{27}$$

Expando y reordeno los factores por potencias de x

$$\frac{(\sigma_1^2+\sigma_2^2)x^2-(2\mu_1\sigma_2^2+2\mu_2\sigma_1^2)x+(\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2} \tag{28}$$

Divido al numerador y el denominador por el factor de x^2

$$\frac{x^2-2\frac{(\mu_1\sigma_2^2+\mu_2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)}x+\frac{(\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)}}{2\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}} \tag{29}$$

Esta ecuación es cuadrática en x, y por lo tanto es proporcional a una función de densidad gausiana con desvío

$$\sigma_\times = \sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \tag{30}$$

y media

$$\mu_\times = \frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)}{(\sigma_1^2 + \sigma_2^2)} \tag{31}$$

Dado que un término $\varepsilon = 0$ puede ser agregado para completar el cuadrado en θ, esta prueba es suficiente cuando no se necesita una normalización. Sea,

$$\varepsilon = \frac{\mu_\times^2 - \mu_\times^2}{2\sigma_\times^2} = 0 \tag{32}$$

Al agregar este término a θ tenemos

$$\theta = \frac{x^2 - 2\mu_\times x + \mu_\times^2}{2\sigma_\times^2} + \underbrace{\frac{\frac{(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)}{(\sigma_1^2 + \sigma_2^2)} - \mu_\times^2}{2\sigma_\times^2}}_{\varphi} \tag{33}$$

Reorganizando el término φ

$$\begin{aligned}
\varphi &= \frac{\frac{(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)}{(\sigma_1^2 + \sigma_2^2)} - \left(\frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)}{(\sigma_1^2 + \sigma_2^2)}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \\
&= \frac{(\sigma_1^2 + \sigma_2^2)(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2) - (\mu_1\sigma_2^2 + \mu_2\sigma_1^2)^2}{\sigma_1^2 + \sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} \\
&= \frac{(\mu_1^2\sigma_1^2\sigma_2^2 + \cancel{\mu_2^2\sigma_1^4} + \cancel{\mu_1^2\sigma_2^4} + \mu_2^2\sigma_1^2\sigma_2^2) - (\cancel{\mu_1^2\sigma_2^4} + 2\mu_1\mu_2\sigma_1^2\sigma_2^2 + \cancel{\mu_2^2\sigma_1^4})}{\sigma_1^2 + \sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} \\
&= \frac{(\sigma_1^2\sigma_2^2)(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{\sigma_1^2 + \sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} = \frac{\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2}{2(\sigma_1^2 + \sigma_2^2)} = \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}
\end{aligned} \tag{34}$$

Luego,

$$\theta = \frac{(x - \mu_\times)^2}{2\sigma_\times^2} + \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \tag{35}$$

Colocando esta forma de θ en su lugar

$$\begin{aligned}
N(x; y, \beta^2)N(x; \mu, \sigma^2) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\underbrace{\left(\frac{(x - \mu_\times)^2}{2\sigma_\times^2} + \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right)}_{\theta}\right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{(x - \mu_\times)^2}{2\sigma_\times^2}\right)\exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right)
\end{aligned} \tag{36}$$

Multiplicando por $\sigma_\times\sigma_\times^{-1}$

$$\frac{\overbrace{\cancel{\sigma_1\sigma_2}}^{\sigma_\times}}{\sqrt{\sigma_1^2 + \sigma_2^2}}\frac{1}{\sigma_\times}\frac{1}{2\pi\cancel{\sigma_1\sigma_2}}\exp\left(-\frac{(x - \mu_\times)^2}{2\sigma_\times^2}\right)\exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \tag{37}$$

Luego,

$$\frac{1}{\sqrt{2\pi}\sigma_\times}\exp\left(-\frac{(x - \mu_\times)^2}{2\sigma_\times^2}\right)\frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}}\exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \tag{38}$$

Retonando a la integral

$$
\begin{aligned}
I &= \int N(x;\mu_\times,\sigma_\times^2)\ \overbrace{N(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)}^{\text{Escalar independiente de x}}\ dx \\
&= N(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)\underbrace{\int N(x,\mu_\times,\sigma_\times^2)dx}_{\text{Integra 1}} \\
&= N(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)
\end{aligned}
\tag{39}
$$