Regresión lineal Bayesiana
Taller Argentino de Computación Científica

Gustavo Landfried

28 de noviembre de 2019

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 $\boldsymbol{w}$.

La regresión lineal, entendida de esta forma, está en la base de los algoritmos más importantes del área de aprendizaje automático e inteligencia artificial y 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.

En este taller vamos a implementar y derivar las distribuciónes de probabilidad que surgen de aplicar la reglas de la probabilidad, i.e. la regla de la suma y el producto, al modelo lineales basadas en transformaciones no lineales fijas. Y debido a que aplicar la reglas de la probabilidad conduce siempre a lo que llamamos inferencia Bayesiana, en este taller veremos la versión Bayesiana de la regresión lineal.

2. Enunciado

Elegir y resolver alguno de los siguientes ejercicios:

1. Implementar una selección de modelo de regresión lineal Bayesianos sobre datos simulados en el lenguaje de programación de su preferencia.
   - *a*) Definir la posterior, la verosimilitud y la evidencia de la regresión lineal Baysiana
   - *b*) Simular datos
   - *c*) Ajustar los datos con regresiones polinomiales de grado 0 hasta 9
   - *d*) Seleccionar modelo basado en la evidencia
2. Derivar la distribución de creencias a poteriori y la evidencia de la regresión lineal Bayesiana.

1. Implementación Para quienes quieran implementar la regresión lineal Bayesiana, van a encontrar una explicación general de la regresión lineal y la definición de la distribución a posteriori, la verosimilitud y la evidencia en la sección **??**

2. Derivación Para quienes quieran comprender de dónde sale la definición de la distribución a posteriori y la evidencia del modelo lineal, van a encontrar la definición del modelo probabilísitico, su modelo gráfico, y las propiedades necesarias para derivar las distribuciones condicional en las secciones **??**.

3. Modelo lineal

Para sacar conlcusiones sobre variables no observables a partir de variables observables es necesario contar con un modelo que indique de qué modo estas variables se relacionan entre sí. Si planteamos un modelo "causal" entre las variables, 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.

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,

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

Donde el vector $\boldsymbol{\phi}(\mathbf{x}_1) = (\phi_0(\boldsymbol{x}_1), \phi_1(\boldsymbol{x}_1), \ldots, \phi_{M-1}(\boldsymbol{x}_1))^T$ 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.

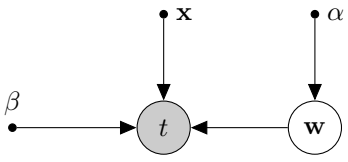
$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{2}$$

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

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

$$P(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) = \mathcal{N}(t|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \beta^{-1}) \tag{3}$$

Planteado en tétminos gráficos,

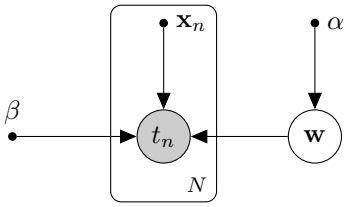


[1] Una pregunta que surge naturalmente es plantearse si las decisiones que tomamos para modelar la relaciones entre variables es la correcta.En este punto no hay que perder de vista que los modelos no son más que representaciones de la realidad, así como los mapas no son el territorio. Esto no significa 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.

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

$$P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \tag{4}$$

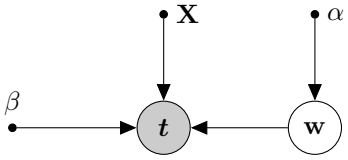
En términos gráficos



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.

$$P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{w}^T\boldsymbol{\Phi}, \beta^{-1}\mathbf{I}) \tag{5}$$

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



$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^T \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^T \end{pmatrix} \tag{6}$$

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.

4. Regresión lineal clásica

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

$$\boldsymbol{w}_{\text{MV}} = \max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) \tag{7}$$

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 (demosración en el anexo).

$$\max_{\boldsymbol{w}} P(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \min_{\boldsymbol{w}} \sum_{i=1}^{n} (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \tag{8}$$

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 encopntrar una configuración $\boldsymbol{w}$ que reduce

aun más la distancia a los datos. Este enfoque conduce a un grave problema conocido como over-fitting (o sobreajuste). Los estrategias comunes para evitar el sobreajuste es agregar términos de regularización ad-hoc a la función de error, evaluación de los parámetros a través de validación cruzada.

5. Regresión lineal Bayesiana

Aquí veremos como 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.

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{9}$$

Due to the choice of a conjugate Gaussian prior, the posterior will also be Gaussian. We can evaluate this distribution by the usal procedure of completing the square in the exponential, and then finding the normalization coefficient using the standard result for normalized Gaussian. The work for deriving the general result is at equation (24)

For simplicity, we consider a zero-mean isotropic Gaussian prior governed by a single precision paramater α so that

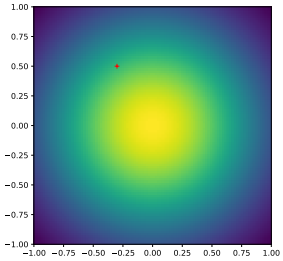
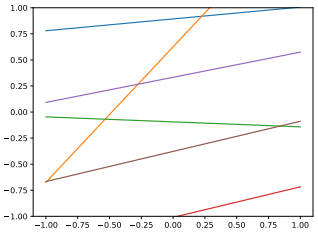
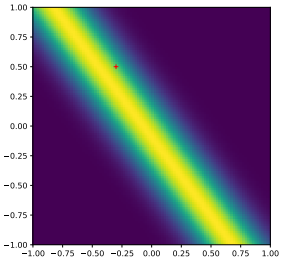
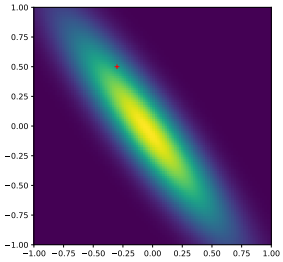
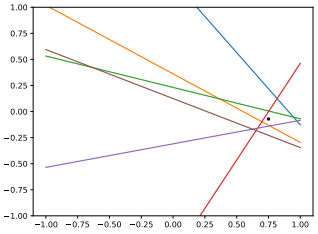
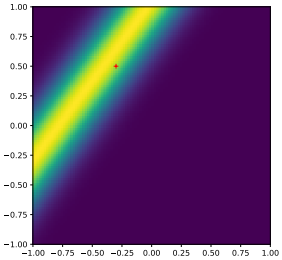
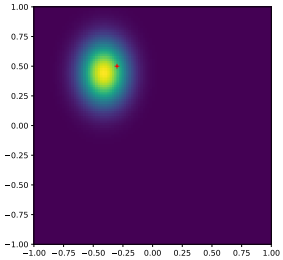
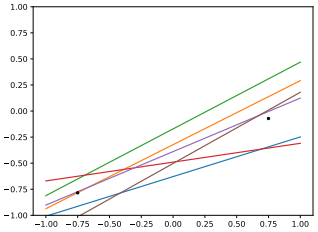
$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{10}$$

then corresponding posterior distribution over $\mathbf{w}$ is then

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t} \tag{11}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi} \tag{12}$$

To obtain an analytical solution we will treat β as a known constant. Note that in supervised learning problems such as regression we are not seeking to model the distribution of the input variables, so we will treat the input $\mathbf{x}$ as a known constant.

Verosimilitud	Priori/Posteriori	Data space
		
		
		

6. Anexo

6.1. Máxima verosimilitud

$$
\begin{aligned}
\log p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w},\beta) &= \sum_{i=1}^{n} \log\, N(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i),\sigma) \\
&= \sum_{i=1}^{n} \log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} = \sum_{i=1}^{n} \log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \log\, e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} \\
&= n\log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \log\, e^{\frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}}} = n\log\, \frac{\sqrt{\beta}}{\sqrt{2\pi}} + \sum_{i=1}^{n} \frac{-(t_i-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\beta^{-1}} \\
&= n\, \log\, \sqrt{\beta} - n\, \log\, \sqrt{2\pi} - \frac{\beta}{2} \sum_{i=1}^{n} (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \\
&\propto - \sum_{i=1}^{n} (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2
\end{aligned}
\tag{13}
$$

6.2. Gaussiana Condicional

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.

Suppose $\mathbf{x}$ is a D-dimensional vector with distributed as a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. And we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$, so that,

$$
\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{14}
$$

We also define corresponding partitions of the mean vector $\boldsymbol{\mu}$ given by

$$
\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{15}
$$

and of the covariance matrix $\boldsymbol{\Sigma}$ given by

$$
\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{16}
$$

Remember that covariance matrix can always be taken to be symmetric, since any antisymmetric component would disapper from the Mahalanobis distance. In many situations, it will be conveninet to work with the precision matrix,

$$
\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \tag{17}
$$

that will be partitioned as follows,

$$
\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{18}
$$

Note that $\boldsymbol{\Lambda}_{aa}$ is not simply given by the inverse of $\boldsymbol{\Sigma}_{aa}$. We shall shortly examine the relationship between them.

From the product rule, the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ can be evaluated from the joint distribution $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ by fixing $\mathbf{x}_b$ to the observed value and normalized the resulting expression.

Instead of performing the normalization explicity, we can consider only the quadratic form of the Mahalanobis distance of thye Gaussian distribution given by (**??**), and then reinstating the normalization at the end.

Making use of the paritioning (14), (15) and (18)

$$\begin{aligned}\Delta^2 =& -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ =& -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\ & -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\end{aligned} \tag{19}$$

Given the partitioning (14), (15), (16) and (18), the **conditional distribution** is,

$$\begin{aligned}p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\end{aligned} \tag{20}$$

Here we shall suppose that we are gioven a marginal Gaussian distribution $p(\mathbf{x})$ and a conditional Gaussian distribution in which $p(\mathbf{y}|\mathbf{x})$ has a mean that is linear function of $\mathbf{x}$.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{21}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}) \tag{22}$$

Given the partitioning (14), (15), (16) and (18), the **Bayes' theorem for Gaussian distribution** is,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\ \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{23}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}[\mathbf{A}^T\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}],\ \boldsymbol{\Sigma}) \tag{24}$$

where,

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{25}$$