Aplicaciones de Machine Learning

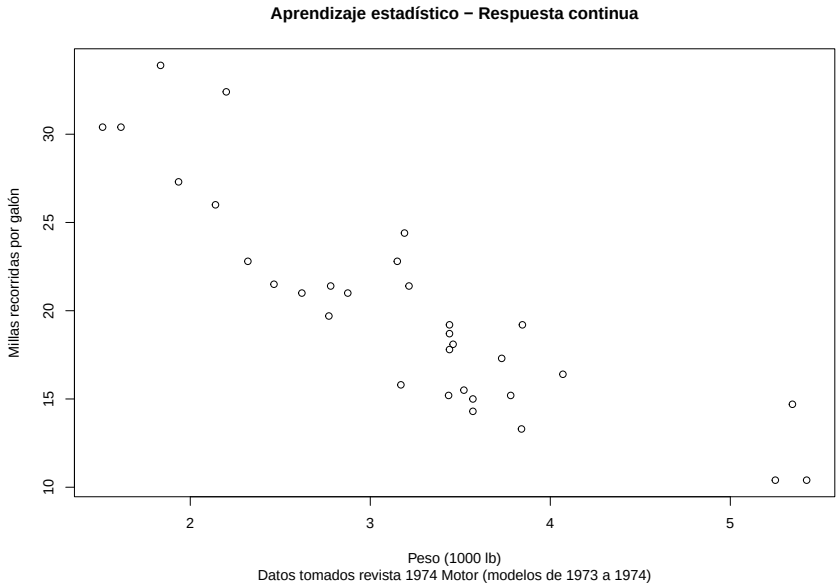
Nicolás López

Primer semestre de 2025

Regresión lineal vs regresión logística

Regresión lineal vs regresión logística

Si visitamos la gráfica de dispersión de los datos de velocidad podemos establecer con claridad una relación entre estas dos variables.



Una relación entre las variables se da de la siguiente forma $Y = \beta_0 + \beta_1 X + \epsilon$.

- R^2 es la **proporción** de la varianza en Y explicada por el regresor X (similar al *odds ratio/ log-odds ratio*)
- F es la **relación** entre la varianza en Y explicada por el regresor X respecto a la que deja de explicar (similar al test de Wald).

En ML buscamos estimar f en $Y = f(X) + \epsilon$, que para RLS resulta ser

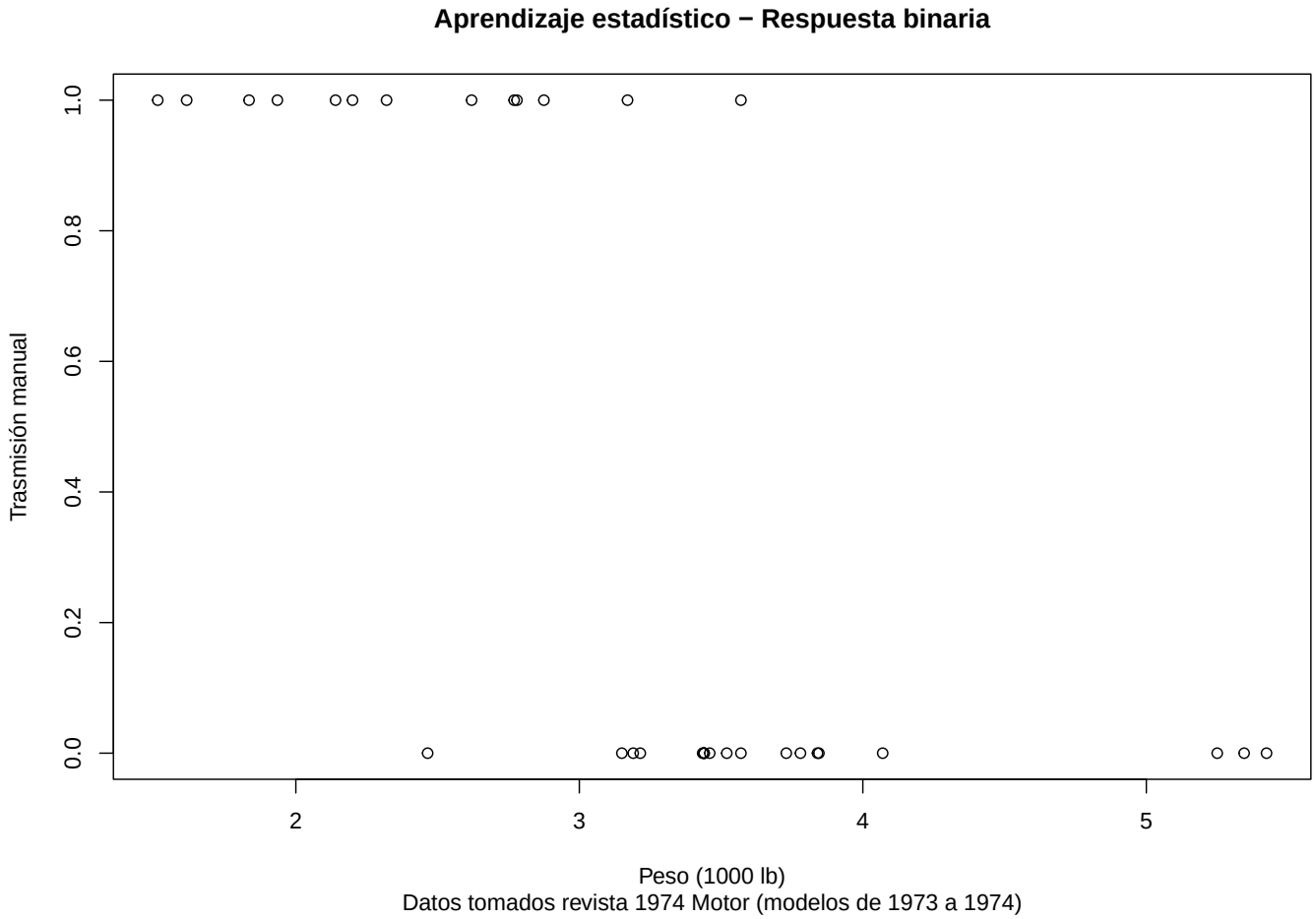
$$f(X) = \beta_0 + \beta_1 X$$

En RLM

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Con X_i covariables discretas o continuas. Cada una con su interpretabilidad bajo el model ajustado.

La diferencia fundamental de la **regresión logística** con RLS/RLM es que nuestra variable respuesta es **binaria**:



En este caso tratamos ahora de un modelo lineal generalizado. Una nueva extensión del método de regresión lineal clásico.

Modelos lineales generalizados

Modelos lineales generalizados

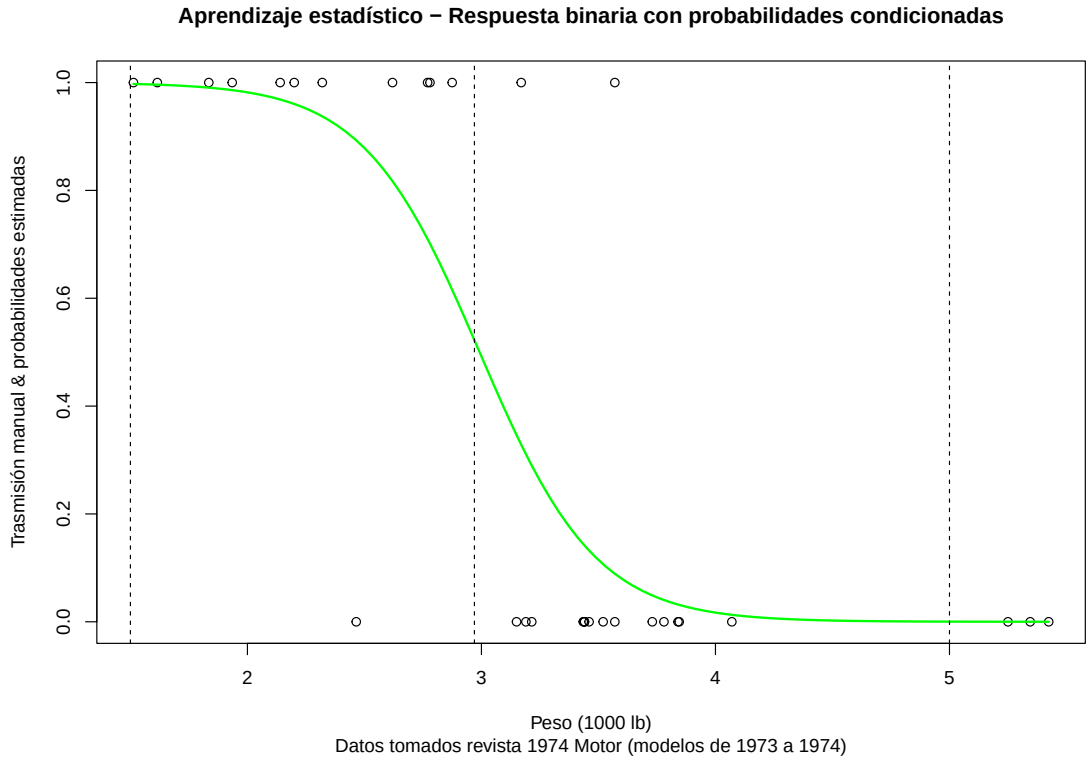
Los modelos lineales generalizados (GLM por sus siglas en inglés) están caracterizados por 3 componentes:

1. Componente aleatorio: variable respuesta Y con una distribución dada.
2. Componente sistemático: covariables para el modelo $X_1,\ldots,X_p$.
3. Función de enlace: Función que relaciona el valor esperado de Y con las covariables $X_1,\ldots,X_p$ de manera lineal.

Diferentes distribuciones de Y dan cabida a una función de enlace especial llamada función de enlace canónico.

Regresión logística

Anteriormente se modelaba el valor esperado de la variable respuesta, siendo esta continua. Nuevamente se modela $\mu(x) = E(Y|X = x)$, sólo que esta vez este valor se encuentra en $[0, 1]$



- Vehículo muy liviano, es altamente probable que sea manual ($y = 1$).
- Vehículo muy pesado, es altamente probable que sea automático ($y = 0$).

Destacando los componentes del GLM se tiene

1. Componente aleatorio: variable respuesta Y con una distribución binomial (éxito y fracaso). En total son 32 carros, entonces $(Y_1, ..., Y_{32})$ son v.a, que se asumen independientes.
2. Componente sistemático: covariables para el modelo $X_1,\ldots,X_p$. En este caso solo tenemos una covariable X igual al peso del vehículo. Una vez observada se relaciona linealmente mediante el predictor lineal:

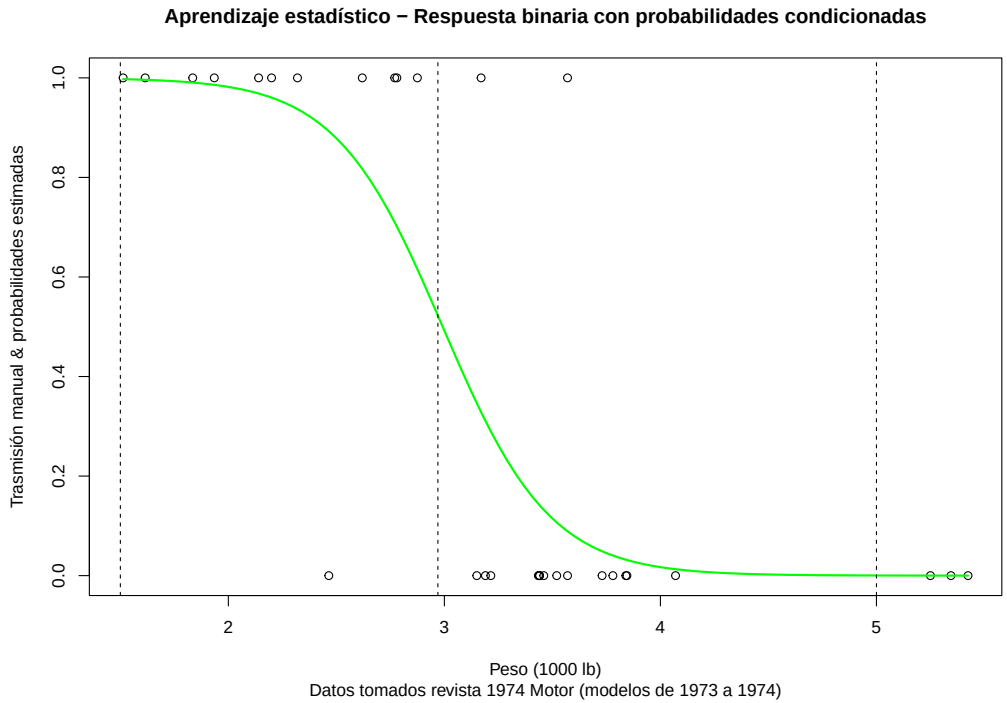
$$\beta_0 + \beta_1 x$$

3. Función de enlace: Función que relaciona el valor esperado de Y con las covariables de manera lineal.

$$g(\mu(x)) = \beta_0 + \beta_1 x$$

Interpretación del modelo

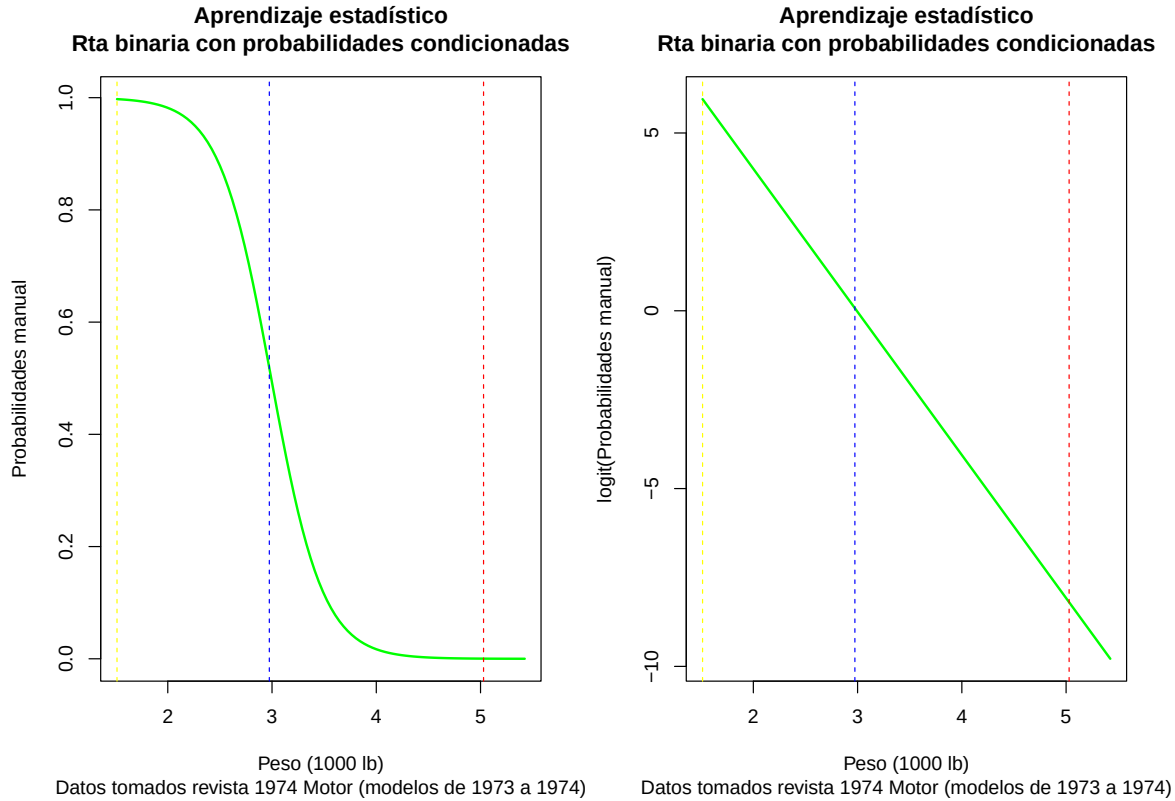
En RLS/RLM, recuerde que nuestra respuesta no está acotada, sin embargo, para el escenario logístico lo está: debe ser una probabilidad (de auto manual) que depende del peso x: $\pi(x)$



Podemos transformar $\pi(x)$ para tener un escenario no acotado como el de RLS/RLM mediante la función de enlace canónico para la distribución bernoulli

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

Con lo cual, $\pi(x)$, la probabilidad de que un carro sea manual dado su peso x, es modelada en el intervalo $(-\infty, +\infty)$.



Y los coeficientes del modelo se presentan en la escala $\text{logit}(\pi(x))$.

Para volver a la escala original (de logit a probabilidad), la función inversa del logit es

$$S(x) = \frac{1}{1 + \exp(x)}$$

La cual es llamada función sigmoide S o logística. Y con esto se tiene que

$$S(\text{logit}(\pi(x))) = \pi(x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

Siendo así la probabilidad de Y modelada a través de $X = x$.

Modelo probabilístico de RLS/RLM

Modelo probabilístico de RLS/RLM

La RLS/RLM es un modelo supervisado que podemos plantear de manera probabilística, con lo cual puede ser estimado mediante máxima verosimilitud en lugar de MCO:

$$P_{\theta=(\beta_0,\beta_1,\sigma)} : y_i \sim N(\mu(x_i), \sigma)$$

Donde

$$\mu(x_i) = E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

Para $i = 1, ..., n$.

Dado P_θ, asumimos que nuestras observaciones siguen dicho modelo, es decir:

$$y_1 \sim N(\beta_0 + \beta_1 x_1, \sigma), ..., y_n \sim N(\beta_0 + \beta_1 x_n, \sigma)$$

Con lo cual la fdp que gobierna el proceso aleatorio generador del dato i-ésimo está dada por:

$$f(y_i|x_i, \theta = (\beta_0, \beta_1, \sigma)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2$$

Para $i = 1, ..., n$. Similar para el caso discreto en la regresión logística.

Con lo cual, la fdp que gobiera el proceso generador de todos los n datos, bajo independencia (recuede que $P(A \cap B) = P(A) \times P(B)$ bajo independencia) es igual a

$$f(y_1, ..., y_n | x_1, ..., x_n, \theta = (\beta_0, \beta_1, \sigma)) = \Pi_i f(y_i | x_i, \theta = (\beta_0, \beta_1, \sigma))$$

Al observar esta fdp conjunta como función de los parámetros, obtenemos la denominada función de verosimilitud del conjunto de datos:

$$L(\theta = (\beta_0, \beta_1, \sigma) | (x_1, y_1), ..., (x_n, y_n)) = \Pi_i f(y_i | x_i, \theta = (\beta_0, \beta_1, \sigma))$$

Y el método de máxima verosimilutud maximiza dicha función:

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}(L(\theta|(x_1, y_1), ..., (x_n, y_n)))$$

Con lo cual se encuentra el $\theta \in \Theta$ más probable para los datos observados. Equivalentemente se puede maximizar la log-verosimilitud, al ser log una función monótona:

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}(l(\theta|(x_1, y_1), ..., (x_n, y_n)))$$

Con

$$l(\theta|(x_1, y_1), ..., (x_n, y_n)) = \sum_i \log\left(f(y_i|x_i, \theta = (\beta_0, \beta_1, \sigma))\right)$$

Para el caso de RLS/RLM se tienen las conocidas soluciones

$$(\hat{\beta}_0, \hat{\beta}_1)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Y

$$\hat{\sigma} = \frac{1}{n}(\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}})$$

Note que para calcular esta verosimilitud, de manera tácita, asume que (x_i, y_i) con $i = 1, ...n$ son observados.

- El algoritmo *EM* (*Expectation Maximization*) es un acercamiento de estimación máximo verosimil en presencia de variables **latentes** (no observables).

Elementos probabilísticos de LDA

Distribución conjunta bajo el modelo LDA de un documento

En este caso también buscamos la distribución conjunta de nuestros datos (como en $f(y_1, ..., y_n)$ anteriormente). Sin embargo, los datos D son ahora las palabras de los M documentos ($D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$).

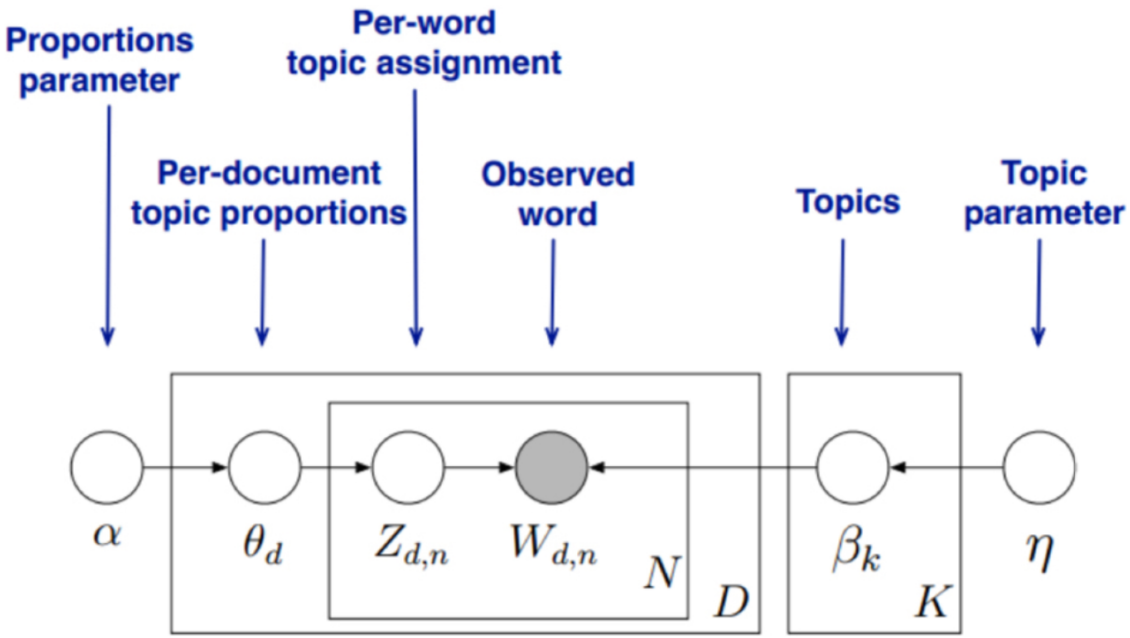


Figure 1: LDA como modelo gráfico (Tomado de Blei)

Paso 1: Proceso Generativo

1. Para cada tema $k \in \{1, ...K\}$
   - Muestreo de $\beta_k \sim \text{Dirichlet}(\eta)$
2. Para cada documento d:
   - Muestreo de $\theta_d \sim \text{Dirichlet}(\alpha)$
   - Para cada palabra n:
     - Elija un tema $z_{dn} \sim \text{Multinomial}(\theta_d)$
     - Elija una palabra $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

Recuerde que Multinomial() es realmente una realización categórica, pero el texto las trata indistintamente.

Paso 2: Factorización de la Distribución Conjunta

Enfocándonos en un documento arbitrario **w** y usando la **regla de la cadena** y dependencias del modelo:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \cdot p(\mathbf{z}|\theta) \cdot p(\mathbf{w}|\mathbf{z}, \beta)$$

Con **z** el vector con la asignación de temas de las N palabras del documento y θ el vector de probabilidades de cada tema.

Con lo cual

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = \underbrace{p(\theta|\alpha)}_{\text{Prior de Dirichlet}} \prod_{n=1}^{N} \underbrace{p(z_n|\theta)}_{\text{Asignación de temas}} \underbrace{p(w_n|z_n, \beta)}_{\text{Generación de palabras}}$$

Dónde

- $p(z_n|\theta) = \theta_{z_n}$, es decir, la probabilidad del tema z_n.
- $p(w_n|z_n, \beta) = \beta_{w_n|z_n}$, es decir, la probabilidad de la palabra w_n para su tema correspondiente z_n.

Paso 3: Desarrollo de Términos Individuales

1. Prior de Dirichlet ($p(\theta|\alpha)$)

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- Γ: Función gamma.
- K: Número de temas.

Paso 3 (Cont.): Desarrollo de Términos

2. Asignación de Temas ($p(\mathbf{z}|\theta)$)

$$p(\mathbf{z}|\theta) = \prod_{n=1}^{N} \theta_{z_n}$$

- θ_{z_n}: Probabilidad del tema z_n en el documento.

3. Emisión de Palabras ($p(\mathbf{w}|\mathbf{z}, \beta)$)

$$p(\mathbf{w}|\mathbf{z}, \beta) = \prod_{n=1}^{N} \beta_{w_n|z_n}$$

- $\beta_{w_n|z_n}$: Probabilidad de la palabra w_n en el tema z_n.

Paso 4: Combinación de Términos

Obteniendo así la verosimilitud para $\theta, \mathbf{z}$ y $\mathbf{w}$ como

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = \underbrace{\frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}}_{\text{Prior de Dirichlet}} \cdot \underbrace{\prod_{n=1}^{N} \theta_{z_n}}_{\text{Temas}} \cdot \underbrace{\prod_{n=1}^{N} \beta_{z_n, w_n}}_{\text{Palabras}}$$

Note que si marginalizamos dos veces tendremos la verosimilitud de un documento

$$p(\mathbf{w}|\alpha, \beta) = \sum_{\theta} \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)$$

Por esto iniciamos con $p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)$.

Y asumiendo independencia para los M documentos ($D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$) se tiene la verosimilitud conjunta del corpus

$$\prod_{d=1}^{M} p(\mathbf{w_d}|\alpha, \beta)$$

En resumen se tiene

- **Tres niveles jerárquicos**:
  1. **Parámetros a nivel de corpus** (α, β):
     - Muestreados una vez para todo el corpus.
  2. **Variables a nivel de documento** (θ_d):
     - Muestreadas una vez por documento.
  3. **Variables a nivel de palabra** (z_{dn}, w_{dn}):
     - Muestreadas para cada palabra en cada documento.
- **Estructura**:
  - $\alpha, \beta \rightarrow \theta_d \rightarrow z_{dn} \rightarrow w_{dn}$.

Como fue destacado anteriormente, una variación del algoritmo EM es utilizado para estimar las probabilidades subyacentes en el modelo, dados los datos.

Modelo de RLS y LDA

Modelo de RLS y LDA

Como podemos notar, el proceso de tópicos es idéntico al mencionado en LDA

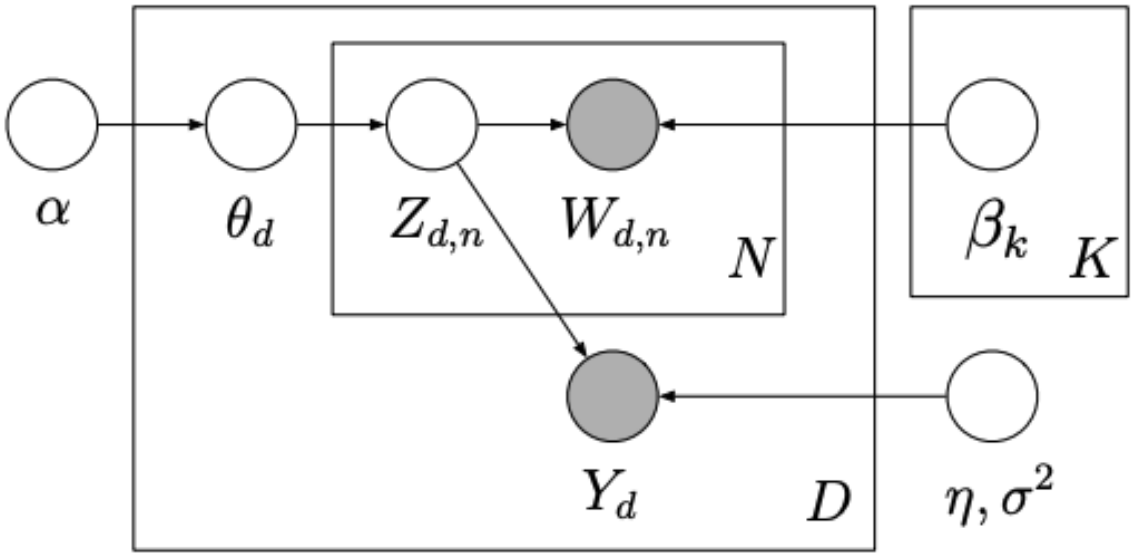


Figure 2: sLDA como modelo gráfico (Tomado de Blei)

Dos elementos importantes aparecen:

1. Y_d Variable respuesta del documento
2. η Coeficientes del predictor lineal junto a la varianza de la distribución normal σ

Proceso Generativo

Para cada documento $d = 1, \ldots, D$:

1. Generar proporciones de temas
2. Generar palabras del documento
3. Calcular estadístico de temas
4. Generar respuesta normal

1. Proporciones de Temas

$$\theta_d \sim \text{Dirichlet}(\alpha)$$

- $\theta_d \in \mathbb{R}^K$: Distribución de temas para el documento d
- $\alpha \in \mathbb{R}^K$: Hiperparámetro de concentración

2. Generación de Palabras

Para cada palabra $n = 1, \ldots, N_d$:

a) Asignación de tema:

$$z_{d,n} \sim \text{Multinomial}(\theta_d)$$

b) Generación de palabra:

$$w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$$

- $\beta_k \in \mathbb{R}^V$: Distribución de palabras para el tema k

3. Estadístico de Temas

$$\bar{z}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n}$$

- $\bar{z}_d \in \mathbb{R}^K$: Frecuencia promedio de temas observada
- Se calcula el promedio empírico de las asignaciones de temas en el documento.
- Este vector $\bar{z}_d$ representa la "frecuencia temática observada" y sirve como predictor en la capa de regresión.

4. Generación de Respuesta

$$y_d \sim \mathcal{N}\left(\eta^\top \bar{z}_d + \mu,\, \sigma^2\right)$$

- $\eta \in \mathbb{R}^K$: Coeficientes de regresión
- $\mu \in \mathbb{R}$: Intercepto
- σ^2: Varianza del error

Modelo de Logístico y LDA

Modelo de Logístico y LDA

En contraste al anterior, la variable respuesta es binaria. Con lo cual el paso 4 es el único que cambia

4. Generación de Respuesta Binomial

$$y_d \sim \text{Binomial}\,(m,\ p_d)$$

donde:

$$\text{logit}(p_d) = \eta^\top \bar{z}_d \quad \Rightarrow \quad p_d = \frac{1}{1 + \exp\left(-(\eta^\top \bar{z}_d)\right)}$$

- m: Número de ensayos (1 para Bernoulli)
- $\eta \in \mathbb{R}^K$: Coeficientes de regresión logística

Referencias

Referencias

1. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Springer.
2. Garet, Witten, Hastie, Tibshirani. Introduction to Statistical Learning with R.
3. Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
4. Blei, D. M., McAuliffe, J. D. Supervised Topic Models. Advances in Neural Information Processing Systems, 20:121–128, 2007.