Clase 1-11

Intro a R y repaso de Proba.

Clase 12: Clasificación

Regresión: Y=variable respuesta, X_i= covariables o variables explicativas.

Estimador no paramétrico de la regresión

Tenemos 2 formas de predecir.

$$Y = \begin{cases} Cuantitativa \longrightarrow Regresion, \\ Cualitativa \longrightarrow Clasificacion. \text{ (Aprendizaje supervisado)} \end{cases}$$

Clasificador: Regla de clasificación que asigna a una muestra una variable respuesta

$$g : X \longrightarrow Y.$$

Error de clasificación medio: $L(g) = P(g(X) \neq Y)$.
Objetivo teorico: Encontrar g que minimice L(g).
Error de clasificación Empírico:
Matemáticamente: $\hat{L}(g) = 1/nsumI_{g(x_i)\neq y_i}$.
En R: $mean(g(X) \neq Y)$.

Regla optima de bayes

- Caso discreto binario:

$$g^{op}(x) = \begin{cases} 1, \text{ si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0, \text{ sino} \end{cases}$$

- Caso continuo: Si $X|Y=0 \sim f_0$, $X|Y=1 \sim f_1$:

$$g^{op}(x) = \begin{cases} 1, \text{ si } f_1(x)P(Y=1) > f_0(x)P(Y=0) \\ 0, \text{ sino} \end{cases}$$

Formas de estimar la distribución condicional de Y dado X

- **k-NN**: Estima a P (Y = 1 — X = x) por la fracción de puntos en N_x={k puntos mas cercanos a x} cuya etiqueta es igual a 1.

$$\hat{P}(Y=1|X=x) = 1/k \sum I_{y_i=1};$$

¿Cómo se elije k?

- **Estimación por núcleos**: considerar un entorno $(x-h, x+h)$ y repetir el procedimiento anterior. Estima a $P(Y=1|X=x)$ por la fracción de puntos en $(x-h, x+h)$ cuya etiqueta es igual a 1:

$$\hat{P}(Y=1|X=x) = \frac{\sum y_i I_{(x-h,x+h}(x_i)}{\sum I_{(x-h,x+h)}(x_i)};$$

¿Cómo se elije h?

Clase 14 Variables Aleatorias

Esperanza es promedio ponderado por las probas de ocurrencia de cada suceso...pero esencialmente es el promedio.

$$E(X) := \sum x f_X(x), \text{ caso discreto;}$$
$$E(X) := \int u f_X(u) du, \text{ caso continuo.}$$

Es la 'mejor' constante que 'resume' o 'aproxima' a nuestra variable aleatoria en el sentido que minimiza la función

$$H(a) = \sum (x_i - a)^2 p_X(x_i) = E((X-a)^2).$$

Es decir, $a = \mu_X$ es la cte tal que $H'(\mu_X) = 0$, i.e. tal que

$$H(\mu_X) = E((X-\mu_X)^2) = Var(X)$$

es mínimo.

La varianza entonces mide cuanto X se aleja de su esperanza, i.e. cuanta "dispersión"tiene.

Desvío standar: $sd(X) = \sqrt{Var(X)}$.

Percentil: el 100p- ésimo percentil de X como el valor x_p que verifica $F_X(x_p) = P(X < x_p) = p$.

Clase 15 y 16: Ley de los Grandes Números

Ley de los Grandes Números (LGN)

Variables iid de $\mathbb{E}[X] = \mu$, $\mathbb{V}[X] = \sigma^2$ entonces

$$\lim_n P(|\overline{X}_n - \mu| > \epsilon) = 0.$$

Es decir el promedio de v.a iid converge en proba a la esperanza.

Esto nos dice que si hay muchos datos el promedio es un buen predictor de la esperanza.

Ej: (X_i) iid X_i F y nos interesa $P(X \in A)$. Definimos $Y_i = I_{X_i \in A} = I_A(X_i)$. Entonces Y_i $B(1,p)$ con $\mathbb{E}(Y_i) = p = P(Y_i = 1) = P(X_i \in A)$, entonces

$$\overline{Y_n} \underset{n}{\to} E(Y_1) = P(X \in A)$$

en proba. Es decir,

$$1/n \sum I_{X_i \in A} \underset{n}{\to} P(X_1 \in A).$$

O sea la frecuencia relativa converge a la probabilidad. Entonces la frecuencia relativa es un buen predictor.

Se define la empírica con el método de plug-in como

$$\hat{F}(t) = \hat{F_n}(t) = \overline{Y_n} = 1/n \sum I_{X_i < t}.$$

En general $1/n \sum g(X_i) \underset{n}{\to} \mathbb{E}(g(X_1))$ (ley del estadista inconsciente).

Estimador: Es una cuenta hecha con la muestra aleatoria, o sea es una v.a. con esperanza y varianza (en general).

Estimación: Es un valor del estimador al usar los datos.

Consistencia: El estimador converge a lo que queremos estimar.

Clase 17: Estimación

Error cuadrático medio:

$$ECM = \mathbb{E}[(\hat{\theta_n} - \theta)^2] = \mathbb{V}(\hat{\theta_n}) + \{\mathbb{E}(\hat{\theta_n} - \theta)\}^2,$$

con $\hat{\theta_n}$ el estimador del parámetro poblacional θ.

Exactitud/Sesgo/Bias:

$$E(\hat{\theta_n}) - \theta.$$

El estimador se dice insesgado si su sesgo vale 0, es decir, si su esperanza coincide con el valor a estimar.

Precisión/Varianza: El estimador se dice preciso si su varianza vale 0, i.e. $\mathbb{V}(\hat{\theta_n})$.

Si $\mathbb{V}(\hat{\theta_n}) \to 0$ y $\mathbb{E}(\hat{\theta_n}) \to \theta$ entonces el ECM tiende a cero, y por tanto el estimador es **consistente**,

$$\hat{\theta_n} \underset{n}{\to} \theta.$$

Ej: Queremos estiamr $\sigma^2 = \mathbb{E}((X - \mu)^2) = \mathbb{V}(X)$ entonces tomamos como estimador a

$$\hat{\sigma}_n = 1/n \sum (X_i - \overline{X})^2.$$

Clase 18 y 19: Método de Máxima Verosimilitud

Objetivo: inferir algo relacionado con el mecanismo (aleatorio) que genera los datos, por ejemplo: ¿cuál es la probabilidad de obtener cara con nuestra moneda?.

Modus Operandi: hacer alguna cuenta con los datos para obtener un valor que se parezca al que queremos estimar: Dados $X_i \sim F$ iid queremos estimar por ej $\mathbb{E}_F(X_1), \mathbb{V}_F(X_1), \mathbb{P}_F(X_1 \leq 50), F$.

- Método plug-in: El procedimiento plug-in propone estimar **mongo**(F) con **mongo**$(\hat{F}_n)$

Modelos paramétricos

Asumimos que la función de distribución F que genera los datos pertenece a una familia conocida, salvo por el valor de cierto parámetro:

$$\mathcal{M} = \{F(\cdot, \theta) :, \theta \in \Theta\},$$

con $\Theta \subset \mathbb{R}^k$ para cierto $k \in \mathbb{N}$.

Caso discreto: $\mathcal{M} = \{P(\cdot, \theta) :, \theta \in \Theta\}$.

Caso continuo: $\mathcal{M} = \{f(\cdot, \theta) :, \theta \in \Theta\}$.

Nos interesa cuán verosímil es que un determinado parámetro haya generado el dato.

Función de verosimilitud

$L(\theta; \mathbf{x})$ mide cuál es la probabilidad de observar nuestra realización $\mathbf{x}$ cuando la probabilidad de cara es θ. Queremos maximizar $L(\theta)$. En general,

- Caso discreto:

$$L(\theta; \mathbf{x}) = P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n) = \prod p_X(x_i).$$

- Caso continuo:

$$L(\theta; \mathbf{x}) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n) = \prod f_X(x_i).$$

Es más fácil maximizar $l(\theta, \mathbf{x}) = \ln L(\theta, \mathbf{x})$, con

- Caso discreto:

$$\ln L(\theta; \mathbf{x}) = \sum \ln p_X(x_i).$$

- Caso continuo:

$$\ln L(\theta; \mathbf{x}) = \sum \ln f_X(x_i).$$

Clase 20: Error de Estimación – Incertidumbre

Varianza y desvío estandar del estimador:

$$\mathbb{V}(\hat{\mu}_n) = \frac{\mathbb{V}(X)}{n},\ se = \sqrt{\mathbb{V}(\hat{\mu}_n)} = \sqrt{\frac{\mathbb{V}(X)}{n}}.$$

Incertidumbre -Error de Estimación

Si $\mathbb{V}(X) = \sigma_0^2$ es un valor conocido, el error de estimación es

$$\text{Incertidumbre-Error de estimacion : } se = \frac{\sigma_0}{\sqrt{n}}.$$

Si $\mathbb{V}(X)$ es desconocido, estimamos se:

$$\text{Incertidumbre-Error de estimacion : } \hat{se}_{obs} = \frac{S_{obs}}{\sqrt{n}},$$

con $S_{obs} = \hat{\sigma}$, x ej sd(datos).

Remuestreo - Bootstrap

Posible aplicación: Asingar incertirudumbre a un percentil muestral.

El método de **Bootstrap** consiste en sacar de nuestro conjunto de datos un dato de manera aleatoria, reponer y repetir. O sea es como simular "nuevos datosçon los datos que ya tenemos. Es hacer indices=sample(lo que sea) y quedarte con datos[indices] (donde en datos repones cada vez que sacás).

Clase 21: Intervalos de confianza

Vamos a pasar de la estimación puntual a la estimación por intervalo.

Vamos a dar un intervalo de valores compatibles con μ.

Diremos que $(a(X_1, \ldots, X_n), b(X_1, \ldots, X_n))$ es un intervalo de confinanza de nivel $1 - \alpha$ para el parámetro θ si

$$P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha.$$

Modelo normal: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$\sigma^2 = \sigma_0^2$ conocido

Sea z_β tal que $P(Z > z_\beta) = \beta$ es decir

$$z_\beta = \phi^{-1}(1 - \beta) = qnorm(1 - beta).$$

En particular, utilizaremos mucho $z_{\frac{\alpha}{2}} = qnorm(1 - alpha/2)$.

Pivot: Cuenta con la muestra y el parámetro de interés, cuya distribución es conocida. Por ej la estandarización de la normal:

$$\hat{\mu}_n = \overline{X}_n, \quad \frac{\hat{\mu}_n - \mu}{\sigma_0/\sqrt{n}} \sim Z, \quad Z \sim \mathcal{N}(0,1).$$

Juntando lo visto tenemos que

$$P(-z_{\alpha/2} < \frac{\hat{\mu}_n - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha \iff P(\hat{\mu}_n - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \mu < \hat{\mu}_n + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}) = 1-\alpha.$$

Es decir,

$$a(X_1,\ldots,X_n) = \hat{\mu}_n - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} = \overline{X}_n - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}};$$
$$b(X_1,\ldots,X_n) = \hat{\mu}_n + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} = \overline{X}_n + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},$$

es un intervalo de confianza de nivel $1-\alpha$ para μ bajo el modelo normal, con varianza σ_0^2 conocida.

σ^2 desconocido

$$\hat{\mu}_n = \overline{X}_n, \quad \frac{\hat{\mu}_n - \mu}{\sigma_0/\sqrt{n}} \sim Z, \quad Z \sim \mathcal{N}(0,1).$$

Entonces,

$$\frac{\hat{\mu}_n - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

con t_{n-1} distribución t-student con $n-1$ grados de libertad. Luego,

$$a(X_1,\ldots,X_n) = \hat{\mu}_n - t_{n-1,\alpha/2}\frac{\sigma_0}{\sqrt{n}} = \overline{X}_n - t_{n-1,\alpha/2}\frac{\sigma_0}{\sqrt{n}};$$
$$b(X_1,\ldots,X_n) = \hat{\mu}_n + t_{n-1,\alpha/2}\frac{\sigma_0}{\sqrt{n}} = \overline{X}_n + t_{n-1,\alpha/2}\frac{\sigma_0}{\sqrt{n}},$$

es un intervalo de confianza de nivel $1-\alpha$ para μ bajo el modelo normal, con varianza σ^2 desconocida, donde en R

$$t_{n-1,\alpha/2} = qt(1 - alpha/2, n-1).$$

Dos Poblaciones: comparación de medias

Muestra 1: $X_1,\ldots,X_n$ iid, $\mathbb{E}(X_i) = \mu_X$ y $\mathbb{V}(X_i) = \sigma_X^2$.
Muestra 2: $Y_1,\ldots,Y_n$ iid, $\mathbb{E}(Y_i) = \mu_Y$ y $\mathbb{V}(Y_i) = \sigma_Y^2$.
Nos interesa saber si las medias de ambas poblaciones son o no iguales.

Uno podría construir un intervalor de confianza para μ_X y uno para μ_Y y si se intersecan concluir que son iguales, pero esto no es ideal.

Un approach mejor es construir un intervalo para $\mu_X - \mu_Y$:

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2},$$

donde $S_p^2 = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}$, y ver si en este intervalo el 0 pertenece para concluir si son iguales.

Mundo asintótico

Diremos que $(a(X_1, \ldots, X_n), b(X_1, \ldots, X_n))$ es un intervalo de confianza de nivel asintótico $1 - \alpha$ para el parámetro θ sii

$$\lim_{n\to\infty} P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha.$$

En este caso lo que podemos asegurar es

$$P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) \approx 1 - \alpha.$$

O sea vamos a tener que aproximar probas así que usaremos TCL.

Por TCL sabemos que

$$\frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{se(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \overset{(a)}{\sim} \mathcal{N}(0, 1).$$

Como $S/\sigma \to 0$ entonces

$$\frac{\sigma}{S}\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \overset{(a)}{\sim} \mathcal{N}(0, 1).$$

Luego

$$\lim_{n\to\infty} P(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{S/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha.$$

Y entonces

$$a(X_1, \ldots, X_n) = \overline{X}_n - z_{\alpha/2}\frac{S}{\sqrt{n}} = \overline{X}_n - z_{\alpha/2}\hat{se};$$
$$b(X_1, \ldots, X_n) = \overline{X}_n + z_{\alpha/2}\frac{S}{\sqrt{n}} = \overline{X}_n + z_{\alpha/2}\hat{se},$$

es un intervalo de confianza de nivel asintótico $1 - \alpha$ para $\mu = \mathbb{E}(X)$. Otra manera de informar esto es $\hat{\mu}(\pm\hat{se})$.

Estimadores asintóticamente normales

En general, si

$$\frac{\hat{\theta}_n - \theta}{\widehat{mongo}} \overset{(a)}{\sim} \mathcal{N}(0,1),$$

entonces,

$$(\hat{\theta}_n - z_{\alpha/2}\widehat{mongo}, \hat{\theta}_n + z_{\alpha/2}\widehat{mongo}),$$

es un intervalo de confianza de nivel asintótico $1-\alpha$ para θ.

Estimadores asintóticamente normales y el método Delta

Si queremos aproximar $f(\theta)$ con f suave entonces por lo anterior tenemos que $\hat{\theta}_n$ asintóticamente normal está dada por

$$\frac{\hat{\theta}_n - \theta}{se} \sim \mathcal{N}(0,1).$$

Haciendo Taylor tenemos

$$\frac{f(\hat{\theta}_n) - f(\theta)}{f'(\theta)se} \sim \mathcal{N}(0,1).$$

Luego

$$\frac{f(\hat{\theta}_n) - f(\theta)}{f'(\hat{\theta})\hat{se}} \sim \mathcal{N}(0,1),$$

y obtenemos que

$$(f(\hat{\theta}_n) - z_{\alpha/2}f'(\hat{\theta})\hat{se}, f(\hat{\theta}_n) + z_{\alpha/2}f'(\hat{\theta})\hat{se}),$$

es un intervalo de confianza de nivel asintótico $1-\alpha$ para $f(\theta)$.

Clase 22 y 23: Estimación No Paramétrica de la Densidad

Clase 24: Regresión no paramétrica

Clase 25: Predicción - Parte II

Clase 26: Revisitando Clasificación

Volvamos un poco al problema de clasificación y a la regla de Bayes de la Clase 12. ¿Cómo podemos obtener las aproximaciones $\hat{f}_0$ y $\hat{f}_1$ de las densidades f_0 y f_1?

Opción 1: Enfoque no paramétrico

Estimamos f_0 y f_1 no paramétricamente con $\hat{f}_{0,h_0}$ y $\hat{f}_{1,h_1}$. Y hacemos un plug-in en la regla $g^{op}(x)$:

$$g^{op}_{h_0,h_1} = \begin{cases} 1, & \text{si } \hat{f}_{1,h_1}(x)\hat{p}_1 > \hat{f}_{0,h_0}(x)\hat{p}_0, \\ 0 & \text{cc .} \end{cases}$$

¿Cómo elegimos k, h o (h_0, h_1)? En cada caso tenemos una $\hat{g}_t$, ¿cómo elegimos el t?

Splitting the data

Test error Podemos repetir el razonamiento que hicimos para regresión: Dado $\hat{r}_{\mathcal{D}_n}$ el estimador construido a partir de $\hat{r}$, consideramos el test error

$$Error_{\mathcal{D}_n} = \mathbb{E}[\{Y_{new} - \hat{r}_{\mathcal{D}_n}(X_{new})\}^2 | \mathcal{D}_n]$$