Clase

Clase 11: taller práctica
12 evaluación

Libros:

Aprendizaje
- supervizado
  - Regresión (Variable continua)
    Predicción
  - clasificación
- no supervisado
- semi-supervizado (Reinforcement learning)
- auto supervisado (LLM)

Regresión: las variables predictoras no tienen xq ser continuas.

Clasificación:

Train, validation, test
80%
Normalizar: $\frac{X-\mu}{\sigma}$

Filtrado de Datos

$\boxed{\text{Underfitting}}$: (sub ajuste) Modelo demasiado simple
$(Y-\hat{Y})$ es grande.

Ej: linea recta para datos con forma de curva.

$\boxed{\text{Sobreajuste}}$

$$E(Y-\hat{Y})^2 = E[f(x)+\varepsilon - \hat{f}(x)]^2$$

$$= E[f(x)-\hat{f}(x)+\varepsilon]^2$$

$$= E[f(x)-\hat{f}(x)^2 + 2(f(x)-\hat{f}(x))\varepsilon + \varepsilon^2]$$

$$= E[f(x)-\hat{f}(x)^2] + 2E[f(x)-\hat{f}(x)] \cdot E(\varepsilon) + E(\varepsilon^2)$$

Sole x q
é um valor
fixo na
sua VA.

$$= f(x)-\hat{f}(x)^2 + 2(f(x)-\hat{f}(x)) \cdot E(\varepsilon) + E(\varepsilon^2)$$

Vale 0
porque
$E(\varepsilon)=0$

$$= f(x)-\hat{f}(x)^2 + \overbrace{E(\varepsilon^2)}^{Var(\varepsilon)}$$

Nota: $Var = E[X - E(X)^2] \Rightarrow Var(\varepsilon) = E(\varepsilon - E(\varepsilon)^2)$

como $E(\varepsilon)=0 \Rightarrow \boxed{Var(\varepsilon) = E(\varepsilon)}$

- Descomposición del error (X es una VA)

$$\text{Error} = (\text{sesgo})^2 + \text{Varianza} + \underline{\text{Error}}\ \underline{\text{irreducible}}$$

$$\text{Err}(X) = E[(Y - \hat{f}(x))^2]$$

$$= \underbrace{f(x) - E(\hat{f}(x))}_{\text{Sesgo}^2}{}^2 + E[\hat{f}(x) - E(\hat{f}(x))^2] + \sigma^2$$

<u>Métricas</u> <u>clasificación</u>

Matriz confusión

	Real: Positivo	Real: Negativo	
Positivo	$\boxed{TP}$	FP	E T II
Negativo	FN.	$\boxed{TN}$	
	Error T I		

Accuracy = $\frac{VP + VN}{VP + VN + FP + FN}$ Verdes / Total = $\frac{23}{32}$

Precision: $\frac{VP}{VP + FP}$

Recall $\frac{VP}{VP + FN}$: Costo de un Falso Negativo

F1_score = $2 \frac{Precision \times Recall}{Precision + Recall}$

. ROC (Receiver Operating Characteristic)

Métricas regresión

$$MSE = \frac{1}{n} \sum \left(y_i - \hat{f}(x_i)\right)^2$$

Validación cruzada

Clase

· Derivar la recta y calcular β_0 y β_1

Regresión múltiple $i, j = 0 \dots n$. $a, b = 0 \dots p$

n y p (p = nº total de variables predictoras)

$$Y_i = Y \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \mathbb{R}^{n\times 1} \quad X = \underset{\mathbb{R}^{n\times p+1}}{} \begin{pmatrix} 1 & X_{11} & X_{12} & X_{1p} \\ 1 & X_{21} & X_{22} & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

X_{ia}

$$\Rightarrow \quad Y = \overbrace{X\beta}^{\hat{Y}} + \varepsilon$$

$$RSS(\beta) = (Y - \hat{Y})^T (Y - \hat{Y})$$

$$= (Y^T - \beta^T X^T)(Y - X\beta)$$

$$= Y^T Y - 2\beta^T X^T Y - \beta^T X^T X \beta$$

Resolvemos

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X^T Y + 2X^T X \beta = 0$$

$$= X^T X \hat{\beta} = X^T Y \quad \Big\} \text{ ecuaciones normales.}$$

$$(X^T) X_{ai} X_{ib} \hat{\beta}_b = (X^T)_{ai} Y_i \quad \Big\}^{\,p+1 \text{ ecuaciones.}}$$

Ejercicio

Si $Q=0$ verifica que.

$\hat{\beta}_0 + \hat{\beta}_a \bar{X}_a = \bar{Y}$ } los promedios estan en el hiperplano.

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$$

Se pueden invertir las matrices siempre que no haya colinealidad (las columnas son CL de otras)

Si $\nexists\, X^T X \Rightarrow$

Coeficiente de determinación

$$R^2 = 1 - \frac{RSS}{TSS}$$

$RSS = \sum (y_i - \hat{y}_i)^2$
Residual

$TSS = \sum (y_i - \langle y \rangle)^2$
total.

Lema: $TSS = RSS + \underbrace{ESS}_{\text{Explain}}$. $ESS = \sum (\hat{y}_i - \langle y \rangle)^2$

Dem:
$$\sum (y_i - \hat{y}_i + \hat{y}_i - \langle y \rangle)^2.$$

qvq: $\sum (y_i - \hat{y}_i)(\hat{y}_i - \langle y \rangle) = 0.$

$$\sum (y - \hat{y}_i)\hat{y}_i - \sum (y_i - \hat{y}_i)\langle y \rangle = 0.$$

Como vale el lema.

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}$$

$$= \frac{1/n \sum (\hat{Y}_i - \langle Y \rangle)^2}{1/n \sum (Y_i - \langle Y \rangle)^2}.$$

R^2: que proporción de la varianza se puede explicar con la varianza del modelo

Clase 3

Regresión logística

Máxima verosimilitud

, dado que $y_i = 1$.

$$\left.\begin{array}{ll} Y_i = 1 & p_i \\ Y_i = 0 & 1-p_i \end{array}\right\} = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\boxed{MLE = \prod p_i^{y_i}(1-p_i)^{1-y_i}}$$ función a maximizar

$$\ln(MLE) = \ln\left(\prod p_i^{y_i}(1-p_i)^{1-y_i}\right)$$

$$= \sum y_i \ln(p_i) + (1-y_i)\ln(1-p_i)$$

$$J(\beta) = -\frac{1}{n}\sum^{n} y_i \ln(p_i) + (1-y_i)\ln(1-p_i)$$

función a MINIMIZAR

$J(\beta)$ es una función que castiga mucho hacer una predicción equivocada (¿por qué?).

Resultado. log-loss para la regresión logística es convexa.

La p_i se relaciona con la β.

Demo de convexidad

$$p_i = \frac{1}{1+e^{-\sum_{a=0}^{p} X_{ia}\beta_a}}$$

$$\frac{\partial J}{\partial \beta_a} = \sum \frac{\partial J}{\partial p_i} \cdot \frac{\partial p_i}{\partial \beta_a}$$

$$\frac{\partial J}{\partial p_i} = -\frac{1}{n}\left(\frac{Y_i}{p_i} - \frac{(1-Y_i)}{1-p_i}\right) = \frac{p_i - Y_i}{n\, p_i(1-p_i)}$$

$$\frac{\partial p_i}{\partial \beta_a} = -\frac{1}{(1+e^{\cdots})^2} \cdot e^{\cdots} \cdot -X_{ia} = p_i(1-p_i)X_{ia}$$

$$\boxed{\frac{\partial J}{\partial \beta_a} = \frac{1}{n} \sum (p_i - y_i)\, x_{ia}.}$$

$$\frac{\partial^2 J}{\partial \beta_a \partial \beta_b} = \frac{1}{n} \sum \frac{\partial p_i}{\partial \beta_b}\, x_{ia}$$

$$= \frac{1}{n} \sum p_i (1 - p_i)\, x_{ia}\, x_{ib}.$$

Quiero ver si la forma cuadrática para un vector arbitrario $v_a\, v_b > 0$

$$\Rightarrow \frac{1}{2} \sum_{a,b} \frac{\partial^2 J}{\partial \beta_a \beta_b} v_a v_b$$

$$= \frac{1}{2n} \sum p_i (1 - p_i) \left(\sum_{a=0}^{p} x_{ia} v_a \right)^2 > 0$$

Porque no lo:

$$J = \frac{1}{2n} \sum (y_i - p_i)^2.$$

$$\frac{\partial^2 J}{\partial \beta_a \partial \beta_b} = p_i(1-p_i)\underbrace{\left[p_i(1-p_i) - (y_i - p_i)(1-2p_i)\right]}^{S_i} x_{ia} x_{ib}.$$

$$S_i\Big|_{y=1} = (1-p_i)(3p_i - 1) < 0 \quad \text{si } p_i < \frac{1}{3} = 0{,}33$$

$$S_i\Big|_{y_i=0} = p_i(2-3p_i) < 0 \quad \text{si } p_i > 2/3.$$

GD

$$\beta_a := \beta_a - \frac{\alpha}{n} \sum (p_i - y_i)\, x_{ia}$$

α = learning-rate

SGD

$$\beta_a := \beta_a - \alpha (p_i - y_i)\, x_{ia}$$

ROC - AUC

		Actual	
		IS obese	Not obese
Predicted	IS obese		
	Not obese.		

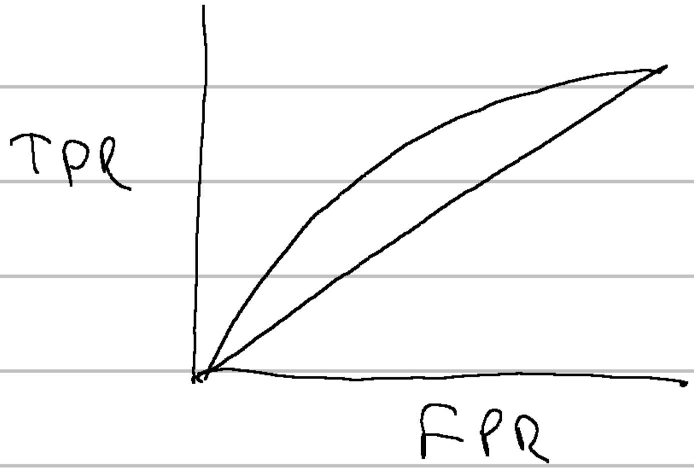
De los casos que importan cuanto detecto.

Sensitivity (Recall) $\frac{TP}{TP+FN}$ $\frac{\text{Pronosticado V}}{\text{TODOS V.}}$

Precision $= \frac{TP}{TP+FP.}$ calidad de la predicho. (Que tan bien predijo cuando acertó)

Accuracy: $\frac{TP+TN}{TOTAL}$ (engañoso)

ROC: Muestra el rendimiento de un clasificador. Para TODOS los umbrales de clasificación



$\boxed{TPR}$ True positive rate (Tasa de verdadera positivos).

Recall

$\boxed{VP / (VP + FN)}$ $TP / (TP + FN)$ = Recall

De todos los que eran positivo (VP+FN) cuantos fueron capaces de predecir.

> Es importante cuando el costo de un FN es alto, eg. detección de cancer.

$\boxed{FPR} = \dfrac{FP}{FP + TN}$ ⇒ • decisiones incorrectas de tratamiento.

Close