

Avaliação de Classificadores

Ricardo Prudêncio

Tópicos

- Métricas
 - Como as predições devem ser avaliadas
- Metodologia de Experimento
 - Como replicar a avaliação para garantir robustez

Matriz de Confusão

Classificação Binária

		PREDITO	
		Classe A	Classe B
VERDADEIRO	Classe A	VP	FN
	Classe B	FP	VN

Classe Positiva =

VP: Verdadeiros Positivos +
FN: Falsos Negativos

Classe Negativa =

VN: Verdadeiros Negativos +
FP: Falsos Positivos

Métricas

- Acurácia:

= $(VP + VN)$ dividido por N ,

onde N é o número total de exemplos

	P^{\wedge}	N^{\wedge}
P	52	6
N	18	44

$$Acc = (52 + 44) / 120 = 0.8$$

maior que o acerto majoritário
(62/120)

Métricas

- Precisão para classe positiva:

= VP dividido por VP + FP

i.e., quanto eu acerto quando dou uma resposta positiva

	P [^]	N [^]
P	52	6
N	18	44

$$\text{Precision} = 52 / (52 + 18) = 0.74$$

Métricas

- Precisão para classe negativa

= VN dividido por VN + FN

i.e., quanto eu acerto quando dou uma resposta negativa

	P [^]	N [^]
P	52	6
N	18	44

$$\text{Precision} = 44 / (44 + 6) = 0.88$$

obs.: nesse exemplo, o valor das
predições negativas é maior

Métricas

- True Positive Rate (Recall, Sensibilidade)

= VP dividido por VP + FN

i.e., quanto eu consigo identificar a classe positiva

	P [^]	N [^]
P	52	6
N	18	44

$$\text{TPR} = 52 / (52 + 6) = 0.89$$

Métricas

- True Negative Rate (Especificidade)

= VN dividido por VN + FP

i.e., quanto eu consigo identificar a classe negativa

	P [^]	N [^]
P	52	6
N	18	44

$$\text{TNR} = 44 / (44 + 18) = 0.70$$

obs.: nesse exemplo, o número de falsos positivo é alto (TNR = 1-FPR)

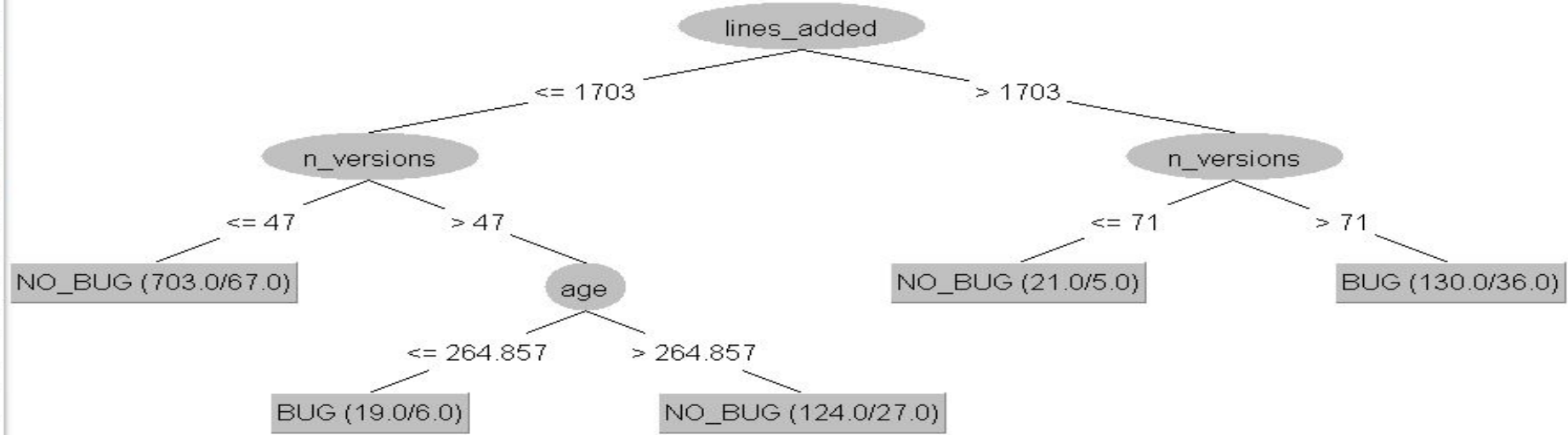
Métricas

- F-Measure (média harmônica de precision e recall)
= $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

OBS.: média harmônica é penalidade quando apenas uma das métricas tem valor alto

Métricas - Análise da CURVA ROC

- Sensibilidade vs Especificidade
 - Muitas vezes conflitantes, em especial quando se usa funções de score
 - Exemplo: porque o colesterol LDL deve ser menor que 130?
Por que não 120?



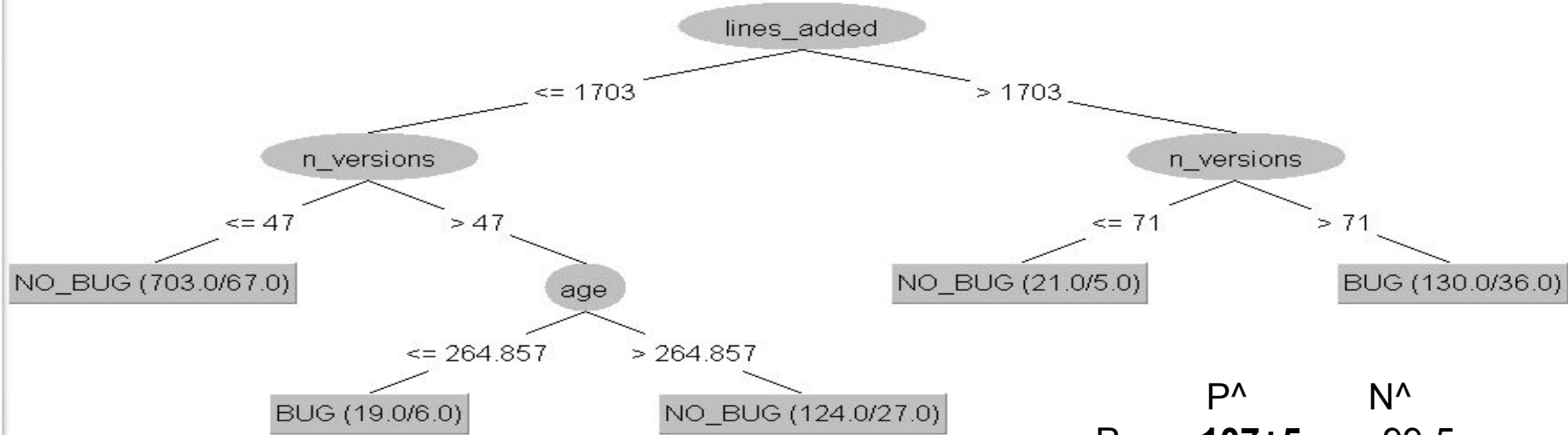
P(Bug|X)

- **BUG** (130/36) : $(130 - 36)/130 = 0.72$
- **BUG** (19/6) : $(19 - 6)/19 = 0.68$
- NOBUG (21/5) : $5/21 = 0.23$
- NOBUG (124/27) : $27/124 = 0.21$
- NOBUG (703/67) : $67/703 = 0.09$

	P [^]	N [^]
P	107	99
N	42	749

TPR = 0.51

FPR = 0.05



Scores

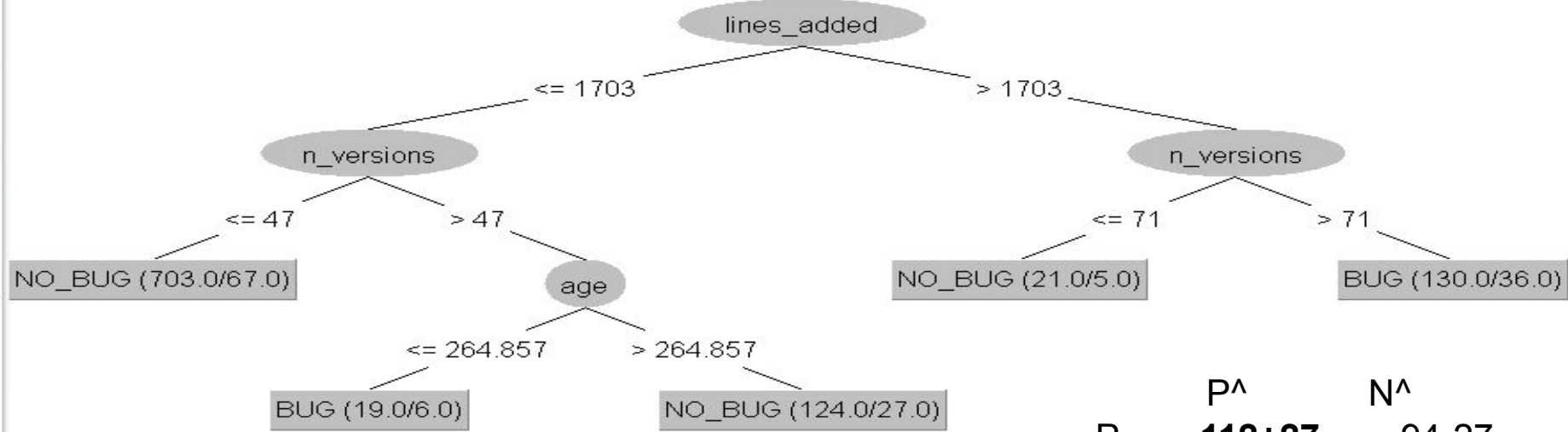
- **BUG** (130/36) : $(130 - 36)/130 = 0.72$
- **BUG** (19/6) : $(19 - 6)/19 = 0.68$
- **BUG** (21/5) : $5/21 = 0.23$
- **NOBUG** (124/27) : $27/124 = 0.21$
- **NOBUG** (703/67) : $67/703 = 0.09$

	P [^]	N [^]
P	107+5	99-5
N	42+16	749-16

	P [^]	N [^]
P	112	94
N	58	733

TPR = 0.54

FPR = 0.07



Scores

- **BUG** (130/36) : $(130 - 36)/130 = 0.72$
- **BUG** (19/6) : $(19 - 6)/19 = 0.68$
- **BUG** (21/5) : $5/21 = 0.23$
- **BUG** (124/27) : $27/124 = 0.21$
- **NOBUG** (703/67) : $67/703 = 0.09$

	P [^]	N [^]
P	112+27	94-27
N	58+97	733-97

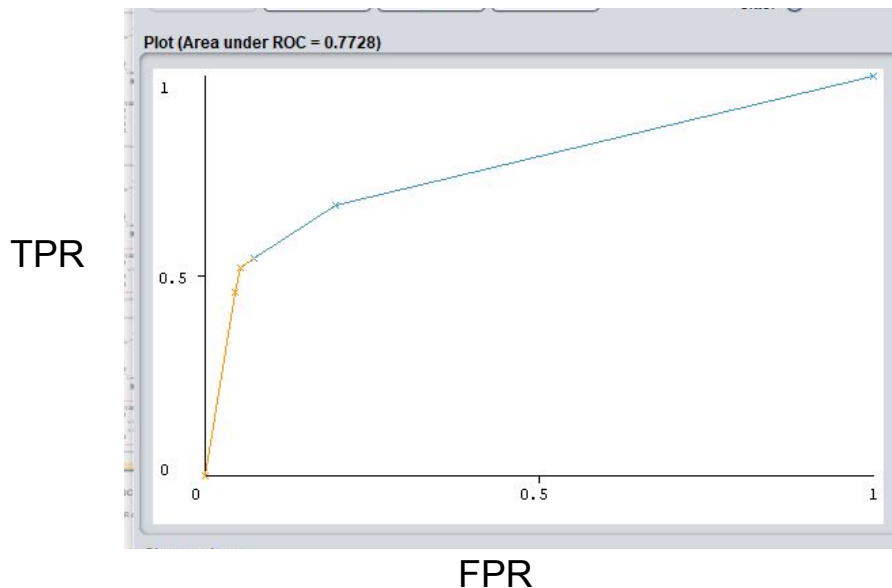
	P [^]	N [^]
P	139	67
N	155	636

TPR = 0.67

FPR = 0.19

Métricas - Curva ROC

- TPR vs FPR considerando diferentes limiares de decisão



Area Under the ROC Curve (AUC) indica qualidade média do modelo considerando diferentes limiares de decisão

Métricas - Curva ROC

- AUC é igual a 1 para um modelo perfeito
- $AUC = 0.5$ pode indicar um modelo aleatório
- AUC = probabilidade de um exemplo positivo escolhido aleatoriamente ter score maior que exemplo negativo escolhido aleatoriamente

Metodologia de Experimentos

- Validação Cruzada K-fold
 - Conjunto de exemplos é dividido em pacotes
 - Por exemplo $K = 10$
 - O algoritmo avaliado é executado K vezes, usando 1 conjunto por vez para teste e o restante para treinamento
 - As métricas de avaliação são calculadas para o conjunto de teste
 - Os resultados médios são calculados entre os conjuntos de teste
 - Opcionalmente a metodologia pode ser repetida usando uma nova divisão dos dados (e.g, 5 k-fold X 2 repetições)

Resumindo

- Existem muitas outras medidas de avaliação
- A melhor medida depende da aplicação
 - Por exemplo, se predições envolverem priorização então uma métrica de qualidade do ranking (i.e. AUC) é adequada
- Métrica deve ser calculada usando uma metodologia de replicação de experimentos