CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Avaliação

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Avaliação de Classificadores

- Procedimento: dividir a base de dados
  - Base de treinamento
  - Base de teste

Acurácia (ou taxa de acerto) do classificador:

$$\frac{\text{nº de acertos (classificações corretas)}}{|\text{base de teste}|}$$

Taxa de erro do classificador:

$$\frac{\text{nº de erros (classificações erradas)}}{|\text{base de teste}|}$$

Técnicas para Avaliação de Classificadores

- ***Hold out***
  - Divisão aleatória (ou preservando a ordem)
  - Base de treinamento (2/3)
  - Base de teste (1/3)
  - Acurácia do classificador é obtida a partir da única base de teste

Random Subsampling

  - Hold out executado **k** vezes
  - Acurácia do classificador é obtida a partir da média das acurácias obtidas nas **k** execuções

Técnicas para Avaliação de Classificadores

- **k-*Fold Cross Validation***
  - Base particionada (aleatoriamente) em k partes (do mesmo tamanho aproximadamente)
  - Treinamento e teste são executados k vezes, e cada execução possui:
  - 1 partição de teste
  - k-1 partições de treinamento
  - Todas as partições são utilizadas, em algum momento, para teste

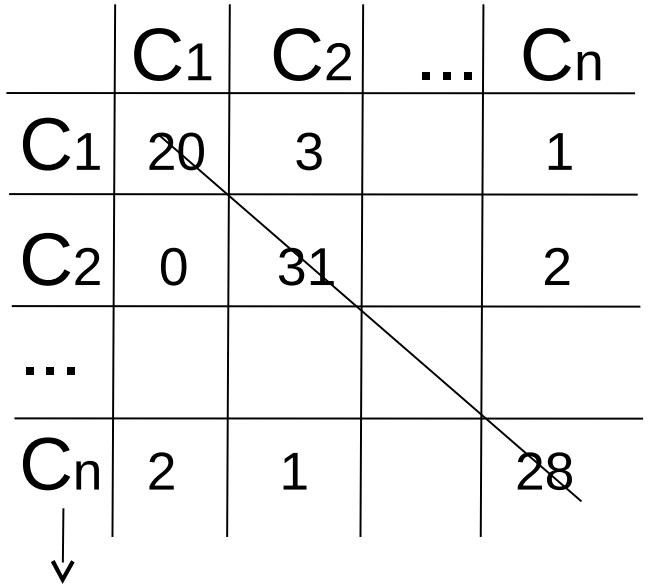
$$\text{acurácia} = \frac{\text{nº total de acertos}}{|\text{base de dados}|}$$

Técnicas para Avaliação de Classificadores

- ***Stratified Cross-Validation***
  - Cada partição utilizada na técnica **k**-Fold Cross Validation deve possuir a mesma distribuição de classes da base original
- ***Leave-one-Out***
  - Mesmo que **k**-*Fold Cross Validation* quando **k** é o número de instância da base de dados

Técnicas para Avaliação de Classificadores

- **Matriz de Confusão**
  - Considere um problema com **n** classes



A célula C_{ij} indica o número de instâncias que foram classificadas na classe C_j e são da classe C_i

Para Classificadores Binários

- Considerando duas classes: spam e não-spam
- **Verdadeiro Positivo** (TP: true positive)
  - Elementos positivos classificados como positivos
  - SPAMs que foram classificados como SPAMs
- **Verdadeiro Negativo** (TN: true negative)
  - Elementos negativos classificados como negativos
  - Não-SPAMs que foram classificados como Não-SPAMs
- **Falso Positivo** (FP: false positive)
  - Elementos negativos classificados como positivos
  - Não-SPAMs que foram classificados como SPAMs
- **Falso Negativo** (FN: false negative)
  - Elementos positivos classificados como negativos
  - SPAMs que foram classificados como Não-SPAMs

Técnicas para Avaliação de Classificadores

- **Para Classificadores Binários**

	spam	**não-spam**
spam	*TP*	*FN*
não-spam	*FP*	*TN*

→ **resultado do classificador**

↓

classe real

Medidas de Avaliação

- Acurácia − representa a porcentagem de elementos do conjunto de teste que foram corretamente classificados

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precisão (Precision) − representa a proporção de elementos classificados como positivos que realmente são positivos

$$P = \frac{TP}{TP + FP}$$

Medidas de Avaliação

- Cobertura (*Recall*) − representa a proporção de elementos positivos que foram classificados como positivos

$$R = \frac{TP}{TP + FN}$$

- F-*Score* ou F-*Measure* − representa a média harmônica entre Precision e Recall

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

Exemplo: matriz de confusão

		Classe Prevista			
		A	**B**	**C**	**Total**
Classe Real	**A**	28	7	3	38
	B	6	32	2	40
	C	5	8	25	38
	Total	39	40	30	116

- Erros = (7+3) + (6+2) + (5+8) = **31**
- Acertos = 28+32+25 = **85** (Diagonal Principal)
- Taxa de erro = 31/116 * 100 = **26,73%**
- Acurácia = 85/116 * 100 = **73,27%**

Exemplo: matriz de confusão

		Classe Prevista			
		A	**B**	**C**	**Total**
Classe Real	**A**	28	7	3	38
	B	6	32	2	40
	C	5	8	25	38
	Total	39	40	30	116

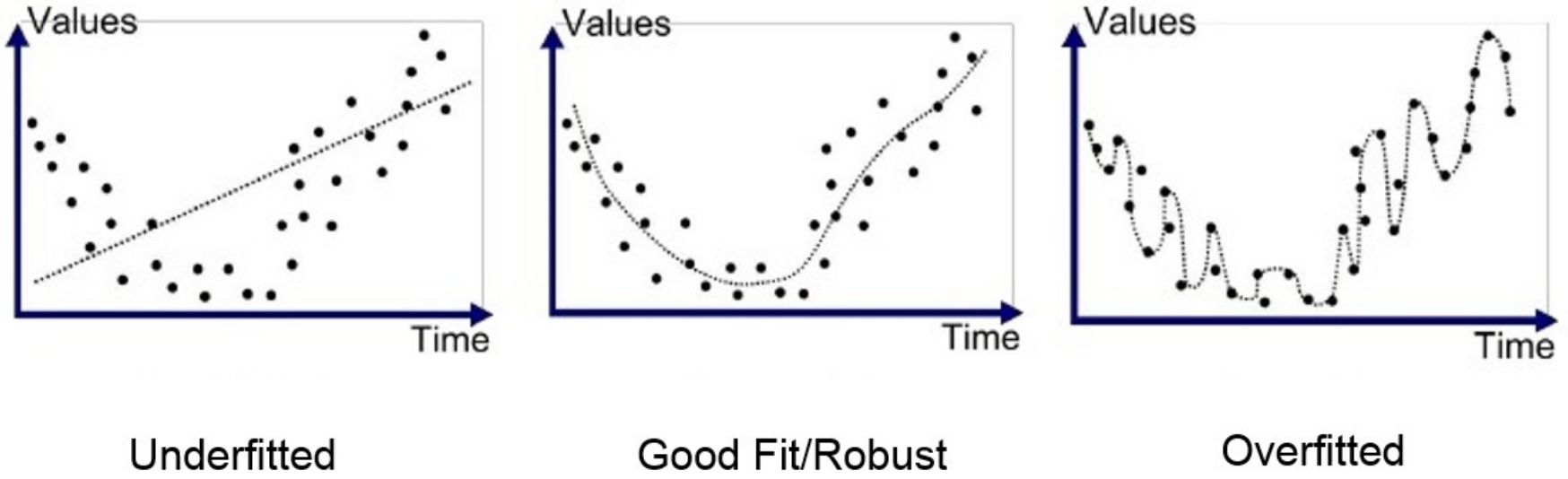
- Análise por classe (*Precision*)
- A = 28/28+6+5 = **0,717**
- B = 32/32+7+8 = **0,680**
- C = 25/25+3+2 = **0,833**

Exemplo: matriz de confusão

		Classe Prevista			
		A	**B**	**C**	**Total**
Classe Real	**A**	28	7	3	38
	B	6	32	2	40
	C	5	8	25	38
	Total	39	40	30	116

- Análise por classe (*Recall*)
- A = 28/28+7+3 = **0,736**
- B = 32/32+6+2 = **0,800**
- C = 25/25+5+8 = **0,657**

Underfitting e *Overfitting*



Underfitted | Good Fit/Robust | Overfitted

- ***Underfitting***
  - Resultados ruins no treino
- ***Overfitting***
  - Resultados bons no treino e ruins no teste

Algumas Considerações

- Cenário/contexto
  - Comparação com outros artigos (estado da arte)
- Número de Classes
  - *Baseline* mínimo
  - 3 classes = 33,33%
- Classe Majoritária
  - Algoritmo *Dummy* (Constant)

CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Avaliação

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com