Aprendizado de Máquinas
Comparando Preditores

Douglas Rodrigues

Universidade Federal Fluminense

Comparando Preditores

- Muitas vezes desejamos saber qual método de construção do preditor performa melhor nos nossos dados.
- Para realizar uma comparação junto, é necessário que todos os métodos tenham utilizado os mesmos parâmetros de pré-treinamento.
- Queremos remover qualquer dependência do desempenho do preditor em relação à amostra selecionada. Para isso, utilizamos métodos de re-amostragem, como *k-fold repetido*.
- Após o treinamento, utilizamos o comando `resamples()` para comparar os resultados.

Comparando Classificadores

- Vamos utilizar o banco de dados `spam`.
- Vamos comparar 4 tipos de treinamento:
  1. **glm**: Modelos Lineares Generalizados.
  2. **svmLinear**: *Support Vector Machine* com *Kernel* Linear.
  3. **rpart**: Árvores de Particionamento e Regressão Recursivas.
  4. **knn**: k-vizinhos próximos.

Comparando Classificadores

- Vamos utilizar *repeated k-folds*, com 3 repetições para $k = 10$.

```
> ctrl <- trainControl(method="repeatedcv", number=10, repeats=3)
```

- Para evitar dependência das amostras treino/teste, vamos fixar `set.seed(100)` antes de cada treinamento.

```
#glm
> {set.seed(100)
> t0 = Sys.time()
> model_glm <- train(type∼., data=spam, method="glm",
trControl=ctrl)
> glm_time<-Sys.time()-t0}
```

-
-
-

Comparando Classificadores

Vamos utilizar alguns comandos que nos auxiliam a comparar os classificadores.

- O comando `resamples()` armazena os dados dos testes, e verifica se os modelos gerados são comparáveis e se utilizam a mesma configuração de `trainControl()`.

```
> results <- resamples(list(GLM=model_glm, SVM=model_svm,
RPART=model_rpart, KNN=model_knn))
```

Comparando Classificadores

- Vamos obter um resumo do resultado dos testes.

  > `summary(results)` Obs: o Coeficiente de Concordância Kappa de Cohen pode ser definido como uma medida de associação usada para descrever e testar o grau de concordância na classificação. É muito importante, principalmente quando há grande discrepância na proporção de diferentes classes na amostra (80% vs 20%, por exemplo).

Comparando Classificadores

Tempos de Processamento

- Método 1

```
> results$timings
> barplot(results$timings$Everything,
names.arg=rownames(results$timings))
```

- Método 2

```
> glm_time
> svm_time
> rpart_time
> knn_time
```

Comparando Classificadores

- Ajustamos o parâmetro dos gráficos.

```
> scales <- list(x=list(relation="free" ),
                 y=list(relation="free" ))
```

- Boxplot comparativo da accuracy e kappa.

```
> bwplot(results, scales=scales)
```

- Densidade da accuracy.

```
> densityplot(results, scales=scales, pch = "|", auto.key=TRUE)
```

- Comportamento de cada fold.

```
> parallelplot(results)
```

Comparando Classificadores

- Comparando comportamento de cada fold em diferentes testes.

```
> xyplot(results, models=c("GLM", "SVM"))
> xyplot(results, models=c("KNN", "RPART"))
```

- Calcular diferença de accuracy/kappa entre modelos, e realizar testes de hipótese para as diferenças.

```
> diffs <- diff(results)
#Resumo e p-valor
> summary(diffs)
```