

Reconhecimento de padrões – aula 14

6 de Janeiro de 2011

Manuel Campagnolo

actualizado em 6 de Janeiro de 2011

Escolha do melhor classificador

Foram analisados diversos classificadores. Muitos deles admitem parâmetros na sua definição, de que derivam classificadores distintos. Será que há classificadores que são, à partida (i.e. independentemente dos dados) melhores do que outros?

Quando não se conhece nada sobre a estrutura dos dados, nenhum classificador é em média, melhor do que outro. Este resultado teórico é conhecido por “no free lunch theorem”.

Nenhum classificador é inerentemente superior

Suponha-se, para simplificar, que há duas classes e que as variáveis são discretas. Então, o valor esperado do erro de classificação é dado por

$$E[\text{Erro}|D] = \sum_{h,F} \sum_{\mathbf{x} \notin D} P(\mathbf{x}) [1 - \delta(F(\mathbf{x}), h(\mathbf{x}))] P(h/D) P(F/D)$$

em que h representa o classificador, D a amostra de treino, F a verdadeira (desconhecida) afectação das observações às classes, e $\delta(x, y) = 1$ se $x = y$ e 0 caso contrário.

Esta igualdade mostra que o valor esperado do erro depende da correspondência entre $P(h/D)$ e $P(F/D)$.

Nenhum classificador é inerentemente superior

O teorema “no free lunch” estabelece então que **se todas as afetações F das observações às classes forem igualmente prováveis**, em média quaisquer dois classificadores têm o mesmo desempenho, i.e.

1. podendo variar D , de tamanho fixo n ,

$$\sum_F \sum_D P(D/F) (E_1[\text{Erro}|F, n] - E_2[\text{Erro}|F, n]) = 0,$$

2. para D fixo,

$$\sum_F (E_1[\text{Erro}|F, D] - E_2[\text{Erro}|F, D]) = 0.$$

Exemplo

Considere uma afectação (desconhecida) F de observações descritas por 3 variáveis binárias a duas classes (indicadas por 1 e -1). A amostra de treino $D = \{000(1), 001(-1), 010(1)\}$ é usada para treinar dois classificadores: h_1 classifica todas as observações em ω_1 excepto as observações que na amostra de treino estão associadas a ω_2 ; h_2 comporta-se de forma exactamente oposta.

1. Qual é o classificador com melhor desempenho sobre a amostra de teste $011(-1), 100(1), 101(-1), 110(1), 111(1)$?
2. Mostre que há 2^5 afectações F distintas que são consistentes com a amostra de treino D .
3. Verifique que para qualquer regra F consistente com D , existe outra regra F' para o qual o desempenho de h_1 e h_2 sobre a amostra de teste resultante é oposto. Em particular, qual é o F' consistente com D , para o qual os resultados da avaliação são opostos aos dos dados na alínea 1?

Princípio da parcimónia

Embora os resultados anteriores indiquem que, sem qualquer conhecimento sobre a regra de decisão correcta, não se deve preferir um classificador a outro, é habitual em classificação preferir classificadores simples a classificadores complexos. Isso pode ser justificado por razões computacionais (um classificador mais simples exige um menor esforço computacional), para reduzir o espaço de procura de classificadores, ou pelo facto de se preferir uma regra que seja mais facilmente interpretável (ver caso das árvores de decisão).

Na prática, classificadores razoavelmente simples tendem a estar associados a menores estimativas de erro. Isso só pode ser justificado pelo facto de se trabalhar em geral com dados cuja estrutura não é arbitrária.

Enviezamento e variância

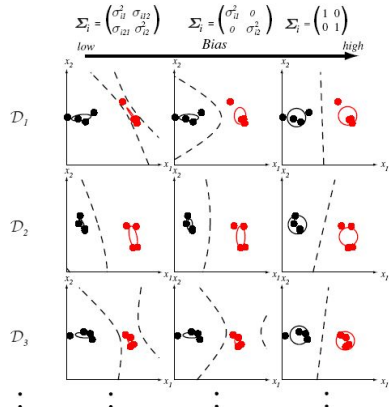
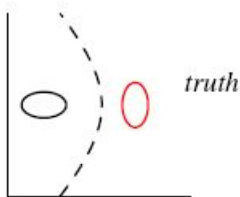
O conceito de enviezamento (“bias”) e variância é muito clara em regressão, em que o erro médio quadrático (MSE), que é o valor esperado do quadrado da diferença entre o verdadeiro valor e a predição do modelo, pode ser decomposto na soma $b^2 + v$, em que b é o enviezamento e v é a variância.

Em classificação não existe uma decomposição semelhante, dado que a variável resposta é discreta. Existem no entanto formas de definir enviezamento e variância como em:

- ▶ J.H. Friedman, On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality, *Data Mining and Knowledge Discovery* 1, 55-77 (1997), ou
- ▶ P. Domingos and G. Hulten. A unified bias-variance decomposition and its applications. In *Proceedings of the 17th International Conference on Machine Learning*, 231-238 (2000).

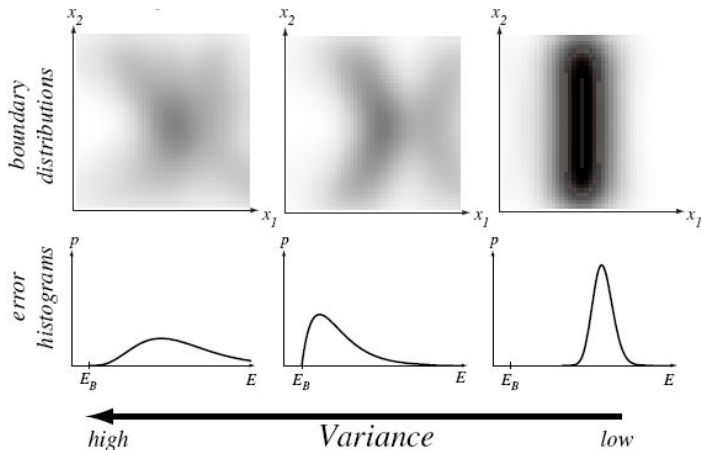
Enviçamento e variância em classificação: exemplo

Considere-se um problema de classificação com duas classes, em que as observações têm uma distribuição normal multivariada, i.e. $P(\mathbf{x}/\omega_i) = N(\boldsymbol{\mu}_i, \Sigma_i)$, em que Σ_i são matrizes diagonais. Na figura abaixo, mostram-se regiões de decisão obtidas usando 3 classificadores paramétricos, para várias amostras de treino.



Enviezamento e variância: continuação do exemplo

Na figura abaixo mostra-se a distribuição das fronteiras de decisão obtidas para grande conjunto de amostras extraídas da população. O classificador mais “rígido” (à direita) tem o maior enviezamento e a menor variância.



Enviezamento e variância

Alguns princípios gerais sobre enviezamento e variância em classificação:

1. Satisfazem uma forma de “lei de conservação”: em geral quando aumento o enviezamento diminui a variância e vice-versa;
2. Em geral, prefere-se um enviezamento moderado a uma elevada variância;
3. Por isso, pode ser preferível usar modelos simples para amostras pequenas, e usar modelos mais flexíveis para amostras maiores;
4. Para o mesmo modelo, o enviezamento e a variância decrescem com o aumento do tamanho das amostras.

Estimar o erro de classificação

Há várias razões para se procurar estimar o erro associado a um classificador para um determinado problema de classificação:

1. Para avaliar a qualidade da classificação obtida sobre novos indivíduos (de classe desconhecida);
2. Para escolher o melhor classificador (avalia-se cada classificador e escolhe-se o que tem o melhor desempenho);
3. Para determinar o melhor valor de um parâmetro para um classificador (por exemplo, o parâmetro de alisamento ou o número de vizinhos num método não paramétrico, um parâmetro do “kernel” numa máquina de vectores de suporte, o número de vértices intermédios numa rede neuronal, ou o número de vértices terminais numa árvore de decisão).

Erro estimado e erro real

O erro de classificação (também designado por erro global) é a probabilidade de o classificador afectar uma observação \mathbf{x} da classe ω_j a uma classe distinta, i.e. $p = P(\alpha(\mathbf{x}) \neq \omega_j)$.

Designa-se por “erro aparente” a proporção de observações da amostra de treino mal classificadas. Esse erro constitui uma estimativa optimista do erro real. Por forma a atenuar esse problema, usam-se estimativas de erro obtidas a partir de observações que não são de treino, designadas por observações de teste ou de validação.

Se dispusermos de uma amostra de teste, com n' observações independentes, o estimador de máxima verosimilhança do erro \hat{p} é a proporção das observações mal-classificadas, i.e. $\hat{p} = k/n'$, em que k é o número de observações de teste mal classificadas.

Exercício. Mostre que k tem distribuição binomial de parâmetros n' e p . Prove que \hat{p} é o estimador de máxima verosimilhança de p .

Intervalos de confiança para o erro dependendo da dimensão da amostra

A não ser que n' seja muito elevado, existe uma grande incerteza sobre o verdadeiro valor de p , dado o valor estimado \hat{p} .

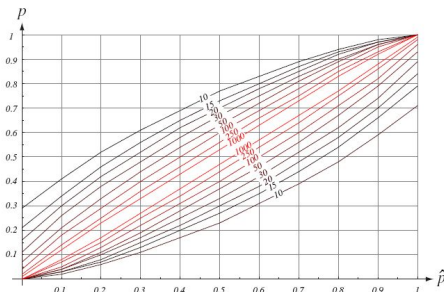


FIGURE 9.10. The 95% confidence intervals for a given estimated error probability \hat{p} can be derived from a binomial distribution of Eq. 3.8. For each value of \hat{p} , the true probability has a 95% chance of lying between the curves marked by the number of test samples n' . The larger the number of test samples, the more precise the estimate of the true probability and hence the smaller the 95% confidence interval. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

“Jackknife” e validação cruzada

A ideia de validação cruzada é formar uma partição da amostra de treino de dimensão n em m conjuntos de observações e usar as observações de $m - 1$ desses conjuntos para treinar o classificador e as observações do conjunto restante para teste. A estimativa de erro de classificação é então a média das m estimativas obtidas dessa forma. Tipicamente, escolhe-se $m = n$ (técnica conhecida como “jackknife”) ou $m = 10$ (validação cruzada propriamente dita).

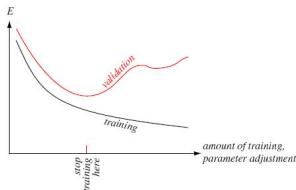


FIGURE 9.9. In validation, the data set \mathcal{D} is split into two parts. The first (e.g., 90% of the patterns) is used as a standard training set for setting free parameters in the classifier model; the other (e.g., 10%) is the validation set and is meant to represent the full generalization task. For most problems, the training error decreases monotonically during training, as shown in black. Typically, the error on the validation set decreases, but then increases, an indication that the classifier may be overfitting the training data. In validation, training or parameter adjustment is stopped at the first minimum of the validation error. In the more general method of cross-validation, the performance is based on multiple independently formed validation sets. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Incerteza associada a estimativas de erro de classificação

Usando a técnica de “jackknife” o estimador de p obtem-se da seguinte forma. $p_{(i)}$ é 0 se a i -ésima observação é classificada correctamente pelo classificador treinado sem essa observação, e é 1 caso contrário. $p_{(\cdot)}$ é a média dos $p_{(i)}$ para $i = 1, \dots, n$ e é o estimador de p . A estimativa da variância desse estimador é dada por $\frac{n-1}{n} \sum_{i=1}^n (p_{(i)} - p_{(\cdot)})^2$.

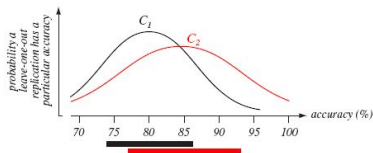


FIGURE 9.11. Jackknife estimation can be used to compare the accuracies of classifiers. The jackknife estimate of classifiers C_1 and C_2 are 80% and 85%, and full widths (twice the square root of the jackknife estimate of the variances) are 12% and 15%, as shown by the bars at the bottom. In this case, traditional hypothesis testing could show that the difference is not statistically significant at some confidence level. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bootstrap

Esta técnica também permite construir amostras de treino e de teste a partir de um conjunto de n observações de classe conhecida. Neste caso, são seleccionados aleatoriamente e **com reposição** n exemplos do conjunto original. Os exemplos seleccionados constituem uma amostra de treino e os restantes exemplos são usados para teste. Podem ser extraídas dessa forma um número arbitrário B de amostras de treino e de amostras de teste correspondentes. O estimativa de p é a média das estimativas obtidas para cada uma dessas amostras.

Em comparação com a o “jackknife” e a validação cruzada, a técnica de “bootstrap” fornece estimativas optimistas do erro (maior enviesamento) mas tem em geral uma menor variância. Aliás, é possível diminuir a variância do estimador aumentando B . É habitual recomendar a técnica de “bootstrap” para amostras relativamente pequenas e a técnica de validação cruzada para amostras maiores.

Algumas limitações das técnicas de estimação

1. Em validação cruzada com m pequeno, os m classificadores podem ser bastante diferentes. É questionável nesse caso usar a estimativa média (sobre os m classificadores) de erro para avaliar o classificador obtido com a totalidade da amostra. Esta limitação é mais crítica para classificadores “instáveis” como as árvores de decisão.
2. Em geral, se a amostra não for muito grande (e.g. > 1000 , ver Isaksson *et al.*, Cross-validation and bootstrapping are unreliable in small sample classification, *Pattern Recognition Letters*, 29 (2008), 1960–1965) existe uma grande incerteza associada ao verdadeiro valor do erro, e por isso os resultados de validação devem ser considerados com algum cuidado.

Aplicação R: estimativas de erro

Diversas funções do R incluem a possibilidade de estimar o erro de classificação, nomeadamente através da técnica de validação cruzada (ver por exemplo `qda` ou `cv.tree`). Há um package dedicado a “bootstrap” (package `boot`)

Para obter estimativas de erro de classificação sobre uma amostra de teste arbitrária, pode construir-se uma **matriz de erro** usando a função `table` e depois calcular \hat{p} como a proporção de observações que não pertencem à diagonal principal da matriz.

```
library(class)
tr <- sample(1:150, 75)
train <- iris[tr, 1:4]; test <- iris[-tr, 1:4]
cl <- iris[tr, 5]; cl.teste=iris[-tr, 5]
iris.knn<-knn(train, test, cl, k=5)
me<-table(iris.knn, cl.teste)
1-sum(diag(me))/sum(me) # estimativa de erro
```