Model evaluation and selection

V. Kitov

Yandex School of Data Analysis

Imperial College London
Department of Physics

January 2015

Table of Contents

Sample usage

- Holdout
- Cross-validation
- Stratified cross-validation

Comments

- Bayes minimum error rate - theoretical lower bound
- Training error rate - optimistically biased
- Test error rate - pessimistically biased (since part of data used for error estimation)

Holdout estimate of error rate distribution

Let e be the probability of making error on previously unseen object.
Probability of observing k errors on test sample of size n:

$$p(k|e,n) = \left(\begin{array}{c} n \\ k \end{array} \right) e^k (1-e)^{n-k}$$

Then

$$p(e|k,n) = \frac{p(e,k|n)}{p(k|n)} = \frac{p(k|e,n)p(e|n)}{\int p(k|n)p(e|n)de}$$

Assuming that $p(e|n) \equiv const$, we obtain

$$p(e|k,n) = \frac{p(k|e,n)}{\int p(k|n)de} \propto e^k (1-e)^{n-k}$$

Since beta-distribution
$Be(x|\alpha,\beta) = [\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))]x^{\alpha-1}(1-x)^{\beta-1}$ it follows that

$$p(e|k,n) \sim Be(k+1, n-k+1)$$

Discriminability vs. reliability

- **Discriminability** measures how well classes are classified
  - Error rate is discriminability measure
- **Reliability** how well class probabilities are estimated
  - Likelihood (y_i is the class of x_i):

$$\prod_{i=1}^{n} \widehat{p}(y_i|x_i)$$

  - Brier score:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\left(\mathbb{I}[x_i \in \omega_c] - \widehat{p}(\omega_c|x_i)\right)^2$$

- Example of good discriminability and poor reliability

Limitation of error rate

- Proportion of errors E
- Proportion of correct classifications $1 - E$
- Average cost

Limitation of error rate

- Proportion of errors E
- Proportion of correct classifications $1 - E$
- Average cost

Limitation

These methods give general performance, without suggesting improvement.

Table of Contents

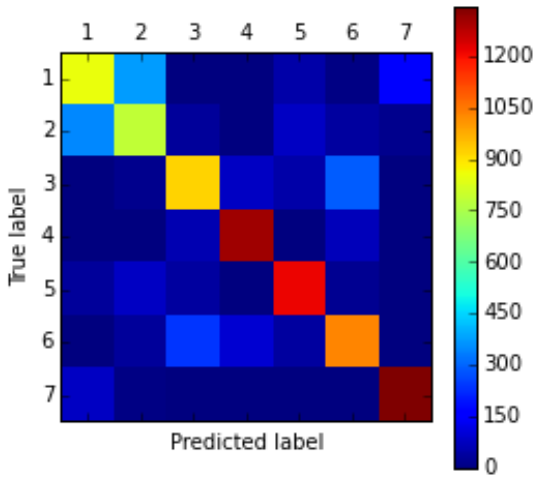
Confusion matrix

$$
\begin{array}{cc} & \text{Estimated classes} \\ & \begin{array}{cccc} 1 & 2 & \cdots & C \end{array} \\ \text{True classes} \begin{array}{c} 1 \\ 2 \\ \vdots \\ C \end{array} & \begin{bmatrix} w_{11} & w_{12} & & \\ w_{21} & w_{22} & & \\ & & \ddots & \\ & & & w_{CC} \end{bmatrix} \end{array}
$$

w_{ij} - number of objects, belonging to ω_i but classified as ω_j.

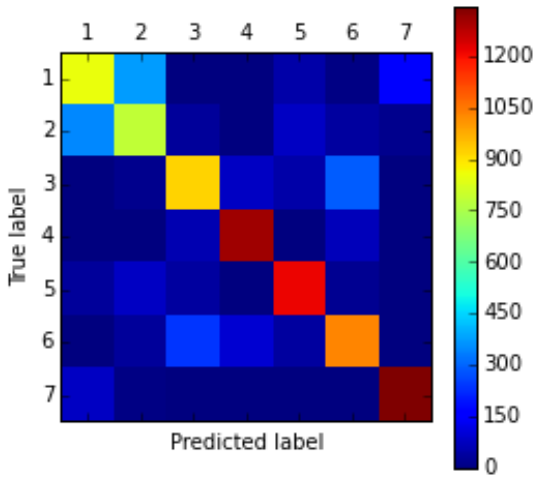
Confusion matrix visualized

Which pair of classes contribute maximum errors?



Confusion matrix visualized

Which pair of classes contribute maximum errors?



Possible solution

Consider hierarchical 2-layer classifier:

- on first layer pairs (ω_1, ω_2) and (ω_3, ω_6) are treated as single classes ω_{12} and ω_{36}
- second layer discriminates individual classes inside grouped classes

2-class case

Confusion matrix:

		Estimated class: +	Estimated class: -
True class	+	True positives	False negatives
	-	False positives	True negatives

2-class case

Confusion matrix:

		Estimated class: +	Estimated class: -
True class	+	True positives	False negatives
	-	False positives	True negatives

Derived performance measures:

Accuracy:	$\frac{TP+TN}{P+N}$	Error rate:	$\frac{FP+FN}{P+N}$
FPR:	$\frac{FP}{N}$	TPR:	$\frac{TP}{P}$
Precision:	$\frac{TP}{TP+FP}$	Recall:	$\frac{TP}{P}$
F-measure:	$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$	F_β-measure:	$\frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{Precision}+\frac{1}{1+\beta^2}\frac{1}{Recall}}$

Table of Contents

Parametrization of predicted class proportions

Bayes minimum risk solution: assign x to ω_1 if

$$\lambda_1 p(\omega_1) p(x|\omega_1) > \lambda_2 p(\omega_2) p(x|\omega_2)$$

This condition is equivalent to

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_2 p(\omega_2)}{\lambda_1 p(\omega_1)} = \mu_1$$

Neyman-Pearson decision: assign x to ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \mu_2$$

where μ_2 is selected so that

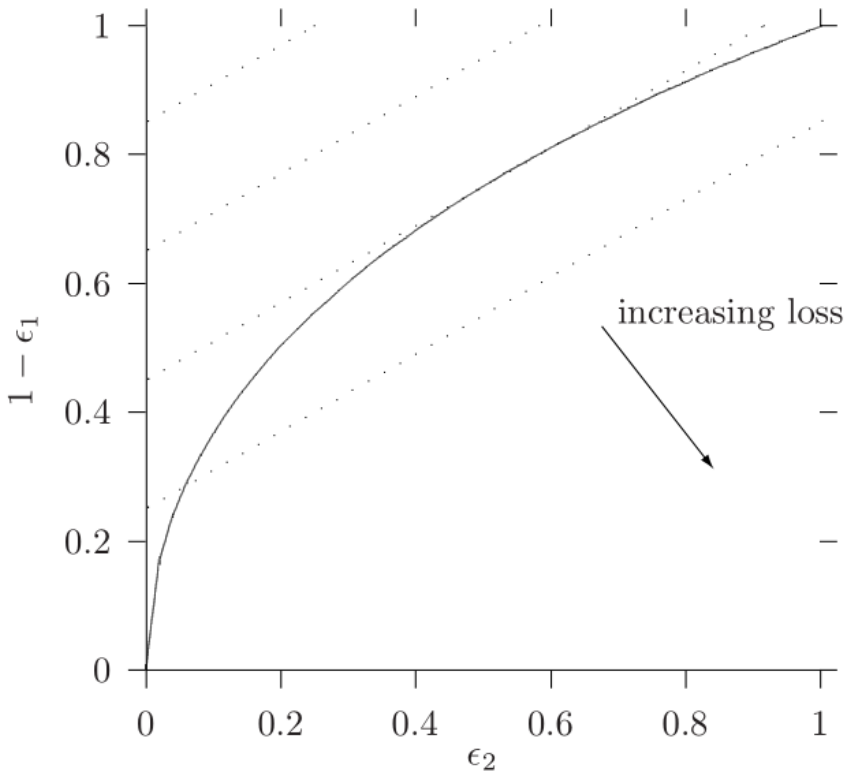
$$\int_{\Omega_1} p(x|\omega_2) dx = \varepsilon_0$$

Discriminant functions example: assign x to ω_1 if $\frac{g_1(x)}{g_2(x)} > \mu$.

ROC curve

As μ increases, the algorithm becomes more inclinable to select class ω_1 (positive class).

- TPR=$1 - \varepsilon_1$ increases
- FPR=ε_2 also increases



ROC properties

- Bayes minimum error, Bayes minimum risk, Newman-Pearson decision rules represent points on the ROC curve.
- Diagonal represents random guessing
- Loss is equal to

$$L = \lambda_2 p(\omega_2)\varepsilon_2 + \lambda_1 p(\omega_1)\varepsilon_1 = \lambda_2 p(\omega_2)\varepsilon_2 - \lambda_1 p(\omega_1)(1-\varepsilon_1) + \lambda_1 p(\omega_1)$$

- At optimality point iso-loss surface is tangent to ROC curve with slope tangent equal to $\frac{\lambda_2 p(\omega_2)}{\lambda_1 p(\omega_1)}$
- Better ROC curves are more concave
- Performance characteristic: area under curve (AUC).
- AUC is also the probability that for random $x_1 \in \omega_1$ and $x_2 \in \omega_2$ if would be true that: $\widehat{p}(\omega_1|x_1) > \widehat{p}(\omega_2|x)$
- Convex hull union of algorithms

Table of Contents

Properties of machine learning algorithms

- Accuracy - achieved on observed data
- Generalization ability (robustness) - expected accuracy decrease on new data
- Online/offline - ability to adapt to new data without full recomputation
- Efficiency
  - computational complexity
    - to train the model
    - to apply the model to new observation
    - to adapt to new information (if online)
  - memory requirements
  - scalability

Properties of machine learning algorithms

- Data constraints
  - assumptions about data distribution
  - flexibility to adapt to data if data assumptions are violated
  - ability to operate well in multidimentional features space
  - possibility to work with discrete/continuous variables or both
- Usability
  - simplicity of maintaince
    - number of user-specified parameters
    - expertise needed to set user-specified parameters
    - simplicity of algorithm logic
  - interpretability of results
- Technical questions:
  - Availability of implicit feature selection
  - Necessity to normalize features before use

Performance assessment caveats

- Quality metric may be different in 3 stages:
  - inside algorithm tuning
  - inside model evaluation
  - real quality in operating conditions
- Population drift: operating conditions may differ from the test set.