ECEN 649 Pattern Recognition

Classification Rules

Ulisses Braga-Neto

ECE Department

Texas A&M University

Sample Data

- In practice, the feature-label distribution F_{XY} is unknown, and so the Bayes classifier is unknown.
- What is available instead is a *sample* from F_{XY}:

$$S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

  where the (X_i, Y_i) are independent and identically distributed (i.i.d.), with $(X_i, Y_i) \sim F_{XY}$.
- The *sample size* n is a deterministic parameter, while

$$n_0 = \sum_{i=1}^{n} I_{Y_i=0} \quad \text{and} \quad n_1 = \sum_{i=1}^{n} I_{Y_i=1}$$

  are binomial *random variables* with parameters $(n, 1-c)$ and (n, c), respectively.

Sample Data - II

- In *separate sampling*, the data are sample from each population separately.

- In this case, the labels $Y_1, \ldots, Y_n$ are not i.i.d. and

$$n_0 = \sum_{i=1}^{n} I_{Y_i=0} \quad \text{and} \quad n_1 = \sum_{i=1}^{n} I_{Y_i=1}$$

  are *deterministic* parameters chosen prior to sampling.

- We will consider this case later in the class. For now, we will concentrate in the case of *mixture sampling* (previous slide).

Classification Rules

- A *classification rule* is a mapping

$$\Psi_n : [R^d \times \{0,1\}]^n \rightarrow \mathcal{C}$$

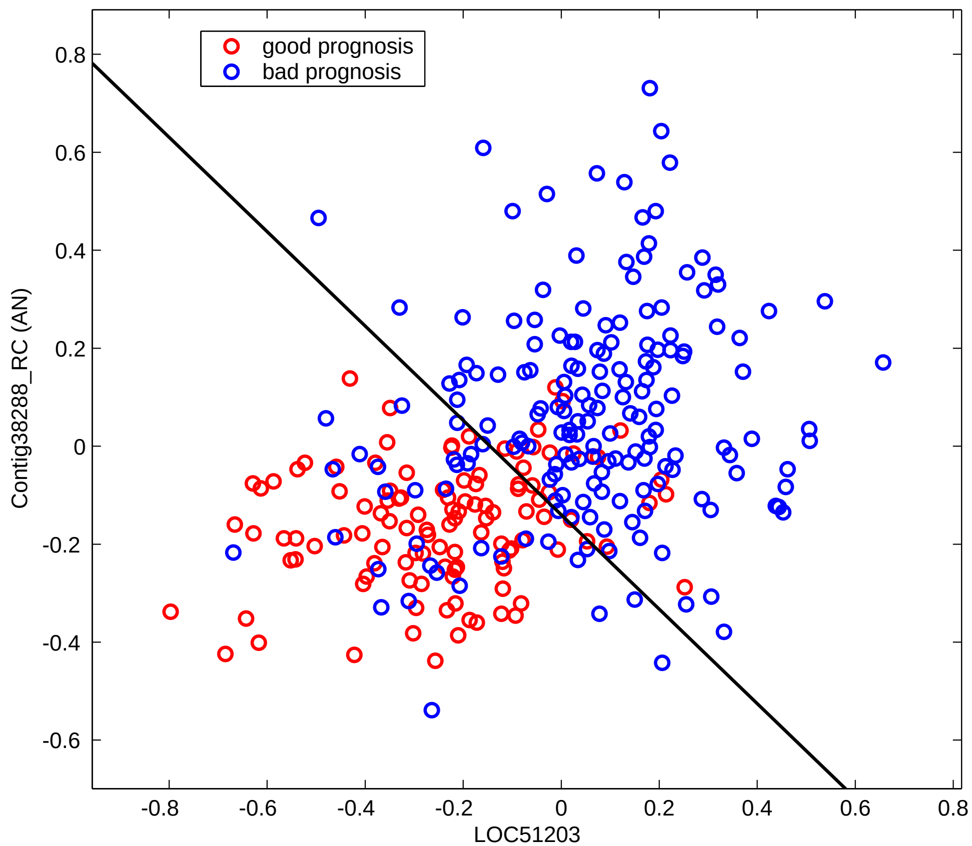
  where $\mathcal{C} = \{\psi \mid \psi : R^d \rightarrow \{0,1\}\}$ is a class containing all classifiers.

- This is simply saying that, given a sample $S_n \in [R^d \times \{0,1\}]^n$, the classification rule Ψ_n produces a designed classifier $\psi_n = \Psi_n(S_n) \in \mathcal{C}$.

- Note that what we have called a classification rule is really a sequence of classification rules depending on n.

Classification Rules - III

Example of designed linear classifier for distinguishing good from bad prognosis among breast cancer patients based on expression of two genes.



Classification Error

- Two kinds of error are of interest here. The first is the familiar classification error of the designed classifier:

$$\epsilon_n = P(\psi_n(X) \neq Y | S_n)$$

This is called the *conditional error* or *true error*.

- The conditional error is a function of the random data S_n, and therefore it is a random variable if the value of S_n is not given. The second kind of error of interest is the expected value of ϵ_n over all sample sets S_n:

$$\mu_n = E[\epsilon_n] = P(\psi_n(X) \neq Y)$$

This is called the *unconditional error* or *expected error*.

Classification Error - II

- The true error ϵ_n is usually the one of most practical interest, since it is the error of the classifier designed on the actual sample data at hand.
- Nevertheless, the expected error $E[\epsilon_n]$ can be of interest precisely because it is data-independent: it is a function only of the classification rule (for a given fixed feature-label distribution F_{XY}).
- Therefore, the expected error can be used to define global properties of classification rules. For example, the most common crriterion for comparing performance of classification rules it to pick the one with smallest expected error $E[\epsilon_n]$ (for a fixed given sample size n and feature-label distribution F_{XY}).

Consistent Classification Rules

- One such global property of classification rules is *consistency*.
- Consistency has to do with the natural requirement that, as the number of samples increases to infinity, classification error should in some sense converge to the Bayes error.
- The classification rule Ψ_n is said to be (weakly) consistent if

$$\epsilon_n \to \epsilon^* \quad \text{in probability}$$

whereas it is said to be strongly consistent if

$$\epsilon_n \to \epsilon^* \quad \text{with probability 1}$$

Consistent Classification Rules - II

- A classification rule Ψ_n is said to be *universally* (strongly) consistent if it is (strongly) consistent for each feature-label distribution F_{XY}.

- While consistency is a property of the classification rule and the feature-label distribution, universal consistency is a property of the classification rule alone.

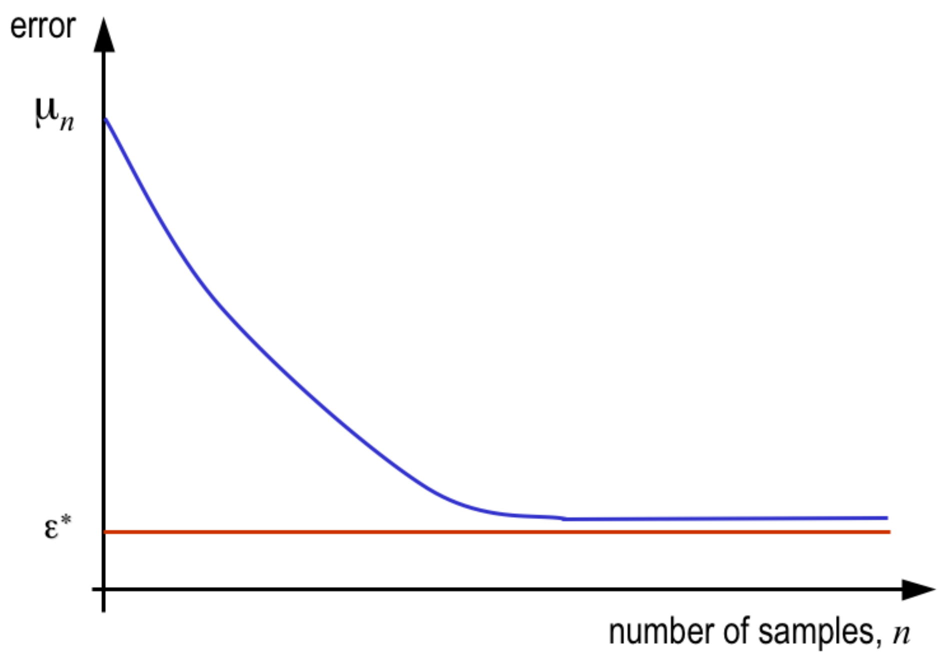
- For example, we will see that the k-nearest neighbor (KNN) classification rule is universally consistent if one lets k increase under a certain rate as n increases, but linear discriminant analysis (LDA) is consistent for certain Gaussian feature-label distributions, while not consistent in general (so it is not universally consistent).

Consistent Classification Rules - III

- The following result relates (weak) consistency to the expected error.
- Theorem: Ψ_n is weakly consistent if and only if

$$E[\epsilon_n] \to \epsilon^*$$

Note that this is ordinary convergence of real numbers.



Consistent Classification Rules - IV

- Therefore, weak consistency means that the classification error ϵ_n is converging to the Bayes error ϵ^* on average, as the data sequence S_n changes with $n \to \infty$.

- Note how much stronger the requirement is for strong consistency: in this case, the classification error ϵ_n converges to the Bayes error ϵ^* for each possible data sequence S_n, as $n \to \infty$, except on a set of data sequences that has probability zero.

Consistent Classification Rules - V

- A word of caution: a non-consistent classification rule may still be useful, in fact, it may be better than a consistent one, in *small-sample* scenarios.
- In the example below, the non-consistent classification rule is better than the consistent one for $n < N_0$.