University of BRISTOL

Computer Science Year 2

Algorithms & Data

Estimation, Regression, Classification

Prof Alin Achim

University of BRISTOL

Two weeks ago, on Thursday …

- Bayesian methods
  - In Bayesian approach the unknown parameter is assumed to be a random variable;
  - They enable prior information about the parameters to be incorporated in the estimation procedure;
  - They do not need to be justified by any asymptotic approximation;
  - Bayesian techniques are based on modelling the uncertainty with respect to the parameter θ through a probability distribution.
  - *Disadvantages*:-
    - A prior distribution must be specified. This presupposes more work and can be subjective
    - Except for some special cases of prior distributions (e.g. Gaussian, Cauchy, exponential, Laplacian), the derivation of the posterior distribution is cumbersome and requires numerical methods.

Previously on DATA …

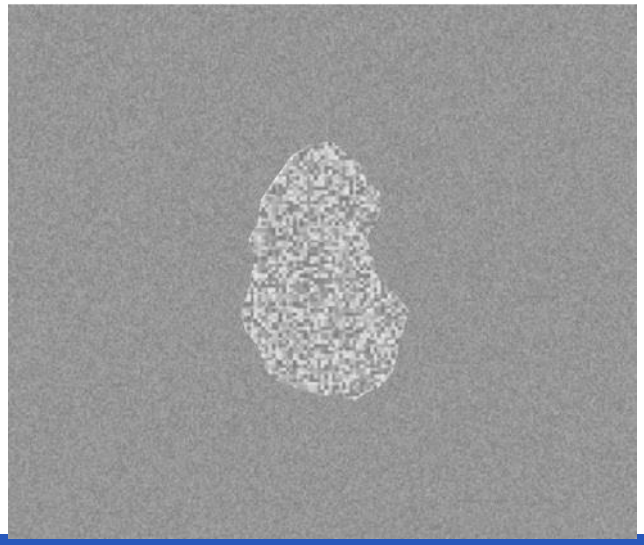
- Maximum Likelihood Estimation (MLE)
  - Based on maximising the likelihood function p(x;θ) which is essentially the probability of the data given the parameters;
  - ML estimators are asymptotically efficient, as the number of observations increase, and the covariance of the estimates tends to CRLB
- Least squares (LS) estimation
  - Minimizes sum of squares between measurements and a model
  - Generally applicable estimator as no assumption is made about the data
- Method of Moments
  - Simplest estimation approach, based on equating sample and population moments
  - Not always leading to good results, especially in small sample sizes

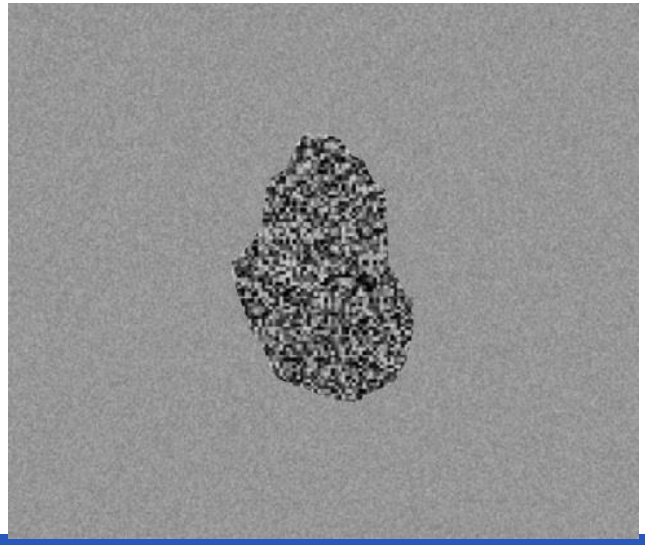
Objectives

- **Classification**
  - Problem definition
  - Bayesian Classifiers
    - Minimising error probability
    - Minimising average risk
  - An Image Analysis Example

Classification: Problem definition

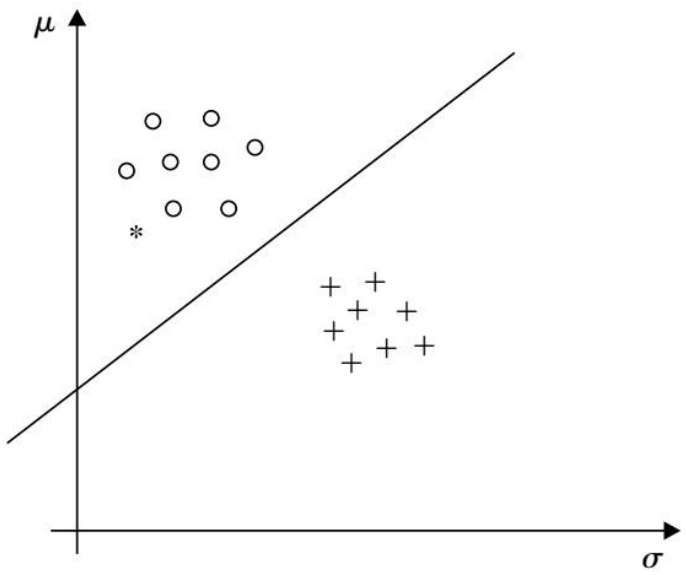
- Consider a simplified case "mimicking" a medical image classification task. The two regions in the images are visually different. We could say that the region in left figure results from a benign lesion, class A, and the other from a malignant one (cancer), class B.
- Assume we have a database of such images with some containing patterns originating in class A, others in B.
- Question is: on receiving a new such image, which class will it be assigned to?





Classification: Problem definition

- First step: determine measurable quantities that make the two regions distinct.
- Plot of the mean value versus the standard deviation for several different images originating from class A (o) and class B (+).
- A straight line separates the two classes.
- On receiving a new image, we measure the mean intensity and standard deviation in the region of interest and we plot the corresponding point (*).



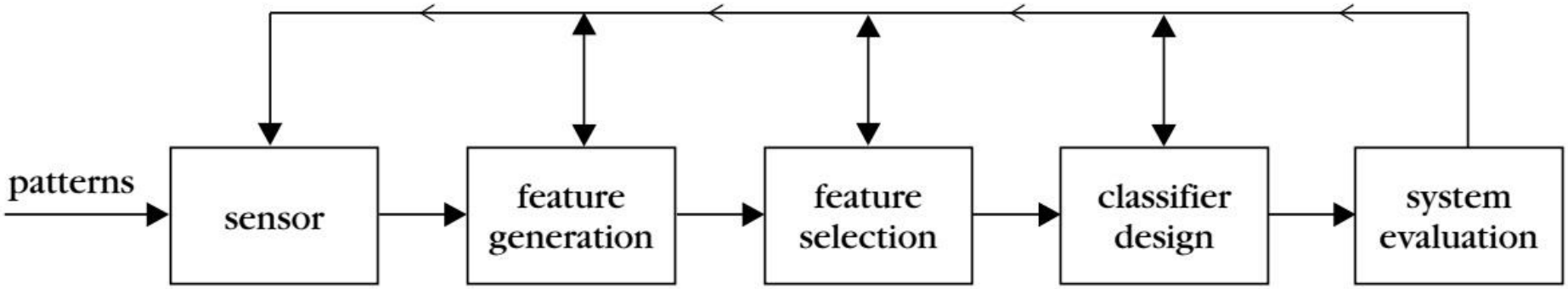
- Mean value μ, and standard deviation σ – **features**
- In general, a number *l* of features is used, and they form a **feature vector**:

$$x = [x_1, x_2, \ldots, x_l]^T$$

- Each feature vector identifies uniquely a single object (pattern)
- Features and feature vectors are treated as random variables and random vectors, respectively.

Image © S. Theodoridis & K. Koutroumbas, *Pattern Recognition*, 4th Edition, 2009.

Standard Classification System



- Questions arising in a classification task:
  - How are features generated? → *Feature generation stage* of design
  - What is the best number l of features to use? → *feature selection stage*
  - How does one design the classifier? → linear or non-linear, what type of non-linearity, and what optimality criterion to use?
    - Supervised, unsupervised, or semi-supervised?
  - How to assess classifier performance? → *system evaluation stage*, e.g. *classification error rate*.

Image © S. Theodoridis & K. Koutroumbas, *Pattern Recognition*, 4th Edition, 2009.

Bayesian Classifiers

- General problem formulation:
  - Feature vector $\boldsymbol{x}, \boldsymbol{x} = [x_1, x_2, \ldots, x_l]$ - represents an unknown pattern
  - Classification task: M classes: $\omega_1, \omega_2, \ldots, \omega_M$
  - M conditional probabilities - $P(\omega_i|\boldsymbol{x}), i = 1, \ldots, M$- *a posteriori* probabilities
    - $P(\omega_i|\boldsymbol{x})$ represents the probability that the unknown pattern belongs to the respective class i , given that the corresponding feature vector takes the value x.
  - Bayesian classifier computes the maximum of these M values or the maximum of an appropriately defined function of them.
- The two-class case:
  - ω_1, ω_2 the two classes in which patterns belong
  - *A priori probabilities* $P(\omega_1|\boldsymbol{x})$ and $P(\omega_2|\boldsymbol{x})$ assumed known
  - *Likelihood functions* $P(\boldsymbol{x}|\omega_i), i = 1,2$ also assumed known

Bayesian Classifiers

- Bayes rule (to compute conditional probabilities):

$$P(\omega_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{P(\boldsymbol{x})}$$

where $P(\boldsymbol{x})$ is the evidence: $P(\boldsymbol{x}) = \sum_{i=1}^{2} P(\boldsymbol{x}|\omega_i)P(\omega_i)$

- *Bayes classification* rule:

If $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_1

If $P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_2

- Alternatively:

If $P(\boldsymbol{x}|\omega_1)P(\omega_1) > P(\boldsymbol{x}|\omega_2)P(\omega_2)$, $\boldsymbol{x}$ is classified to ω_1

If $P(\boldsymbol{x}|\omega_1)P(\omega_1) < P(\boldsymbol{x}|\omega_2)P(\omega_2)$, $\boldsymbol{x}$ is classified to ω_2

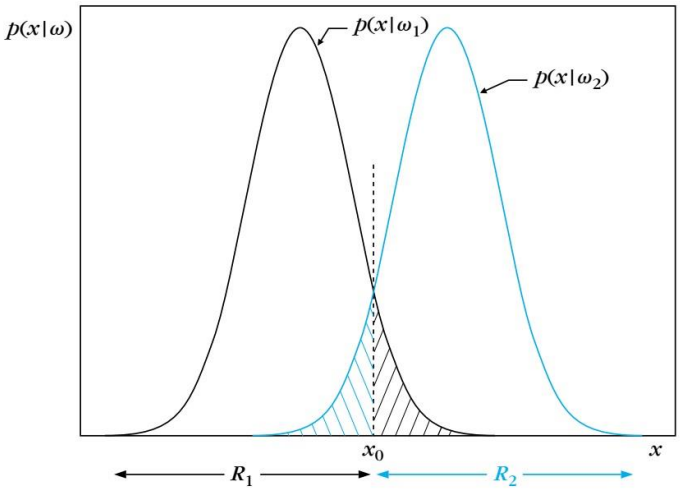
Bayesian Classifiers

- Moreover: If the a priori probabilities are equal ($P(\omega_1) = P(\omega_2)$) then previous rule becomes

$$\text{If } P(\boldsymbol{x}|\omega_1) > P(\boldsymbol{x}|\omega_2), \boldsymbol{x} \text{ is classified to } \omega_1$$

$$\text{If } P(\boldsymbol{x}|\omega_1) < P(\boldsymbol{x}|\omega_2), \boldsymbol{x} \text{ is classified to } \omega_2$$

- Plot shows above situation for a single feature, i.e. $l = 1$
- Dotted line at x_0 is a threshold partitioning the feature space into two regions, R_1 and R_2.
- For all values of x in R_1, the classifier decides ω_1 and for all values in R_2 it decides ω_2.

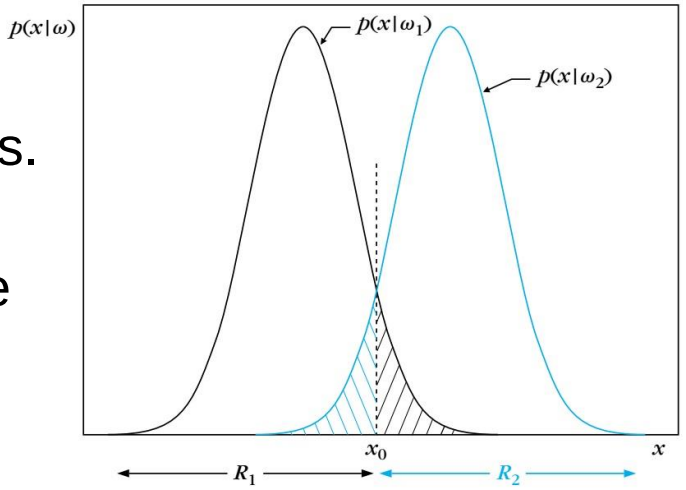


- Probability of decision error: $P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{\infty} p(x|\omega_1)dx$

Image © S. Theodoridis & K. Koutroumbas, *Pattern Recognition*, 4th Edition, 2009.

Classification Error Probability Minimisation

- *Bayesian classifier is optimal with respect to minimising classification error probability.*
- Moving the threshold away from x_0 always increases the shaded area under the curves.
- Formally, an error is made if we decide $\boldsymbol{x} \in R_1$ although it belongs to ω_2 or if we decide $\boldsymbol{x} \in R_2$ although it belongs to ω_1:



$$P_e = P(\boldsymbol{x} \in R_2, \omega_1) + P(\boldsymbol{x} \in R_1, \omega_2)$$

$$P_e = P(\boldsymbol{x} \in R_2|\omega_1)\mathrm{P}(\omega_1) + P(\boldsymbol{x} \in R_1|\omega_2)\mathrm{P}(\omega_2)$$

$$= \mathrm{P}(\omega_1)\int_{R_2} P(\boldsymbol{x}|\omega_1)\,d\boldsymbol{x} + \mathrm{P}(\omega_2)\int_{R_1} P(\boldsymbol{x}|\omega_2)\,d\boldsymbol{x}$$

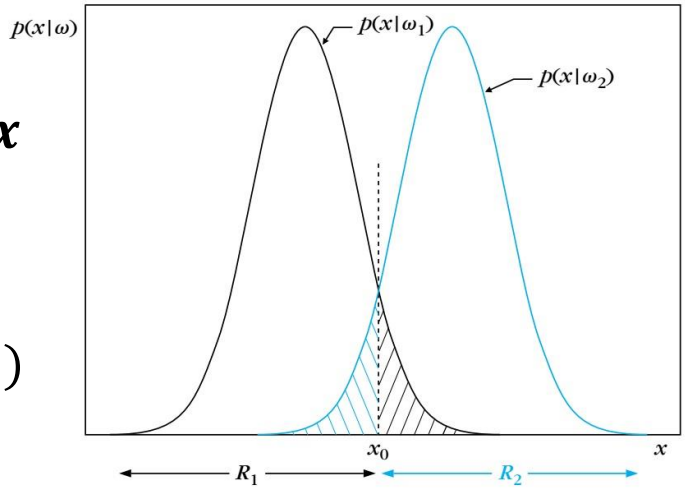
$$= \int_{R_2} P(\omega_1|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_1} P(\omega_2|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$$

$$P(\omega_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{P(\boldsymbol{x})}$$

Classification Error Probability Minimisation

- *Bayesian classifier is optimal with respect to minimising classification error probability.*

$$P_e = \int_{R_2} P(\omega_1|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_1} P(\omega_2|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$$



- Union of R_1 and R_2 cover all the space:

$$\int_{R_1} P(\omega_1|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_2} P(\omega_1|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} = P(\omega_1)$$

Combining above two equations:

$$P_e = P(\omega_1) - \int_{R_1} \big(P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x})\big)P(\boldsymbol{x})\,d\boldsymbol{x}$$

- Hence, error is minimised when R_1 is:

$$R_1 : P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$$

Similarly, R_2 is:

$$R_2 : P(\omega_2|\boldsymbol{x}) > P(\omega_1|\boldsymbol{x})$$

Classification Error Probability Minimisation

- *Bayesian classifier is optimal with respect to minimising classification error probability.*

- Also true for the multiclass case

  - In a classification task with M classes, $\omega_1, \omega_2, \ldots, \omega_M$, an unknown pattern, represented by the feature vector $\boldsymbol{x}$, is assigned to class ω_i if

$$P(\omega_i|\boldsymbol{x}) > P\left(\omega_j\middle|\boldsymbol{x}\right) \forall j \neq i$$

Example

- In a two-class problem with a single feature, x, the pdfs are Gaussians with variances $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively:

$$P(x|\omega_1) = \frac{1}{\sqrt{\pi}} exp(-x^2) \text{ and } P(x|\omega_2) = \frac{1}{\sqrt{\pi}} exp(-(x-1)^2)$$

- If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value, x_0, for minimum error probability.

- The threshold is the point where $P(\omega_1|x) = P(\omega_2|x)$, or (using Bayes theorem) where $P(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2)$.
- Since $P(\omega_1) = P(\omega_2)$, it follows that the threshold will be found where $P(x|\omega_1) = P(x|\omega_2)$.
- Substituting: x_0: $exp(-x^2) = exp(-(x-1)^2)$
- Taking log: x_0: $-x^2 = -(x-1)^2$
- Finally: $x_0 = 1/2$

Minimising the Average Risk

- The classification error probability is not always the best criterion to be adopted for minimization.
- It assigns the same importance to all errors.
- There are cases in which some wrong decisions may have more serious implications than others.
- In such cases it is more appropriate to assign a penalty term to weigh each error.
- By minimising a modified version of the error probability, P_e, i.e.:

$$\mathrm{r} = \lambda_{12}\mathrm{P}(\omega_1)\int_{R_2} P(\boldsymbol{x}|\omega_1)\,d\boldsymbol{x} + \lambda_{21}\mathrm{P}(\omega_2)\int_{R_1} P(\boldsymbol{x}|\omega_2)\,d\boldsymbol{x}$$

a different optimality criterion emerges.

Minimising the Average Risk

- Without proof:

$$l_1 = \lambda_{11} P(\boldsymbol{x}|\omega_1) P(\omega_1) + \lambda_{21} P(\boldsymbol{x}|\omega_2) P(\omega_2)$$
$$l_2 = \lambda_{12} P(\boldsymbol{x}|\omega_1) P(\omega_1) + \lambda_{22} P(\boldsymbol{x}|\omega_2) P(\omega_2)$$

- We assign $\boldsymbol{x}$ to ω_1 if $l_1 < l_2$, i.e.:

$$(\lambda_{21} - \lambda_{22}) P(\boldsymbol{x}|\omega_2) P(\omega_2) < (\lambda_{12} - \lambda_{11}) P(\boldsymbol{x}|\omega_1) P(\omega_1)$$

- Correct decisions are penalised less than wrong ones, so $\lambda_{ij} > \lambda_{ii}$, hence the classification rule becomes:

$$\boldsymbol{x} \in \omega_1 (\omega_2) \text{ if } l_{12} = \frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)\lambda_{21} - \lambda_{22}}{P(\omega_1)\lambda_{12} - \lambda_{11}}$$

- l_{12} is the l*ikelihood ratio* and the above test is the *likelihood ratio test*.

- $L = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$ is called the *loss matrix*.

Minimising the Average Risk – Example

- In a two-class problem with a single feature, x, the pdfs are Gaussians with variances $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively:

$$P(x|\omega_1) = \frac{1}{\sqrt{\pi}} exp(-x^2) \text{ and } P(x|\omega_2) = \frac{1}{\sqrt{\pi}} exp(-(x-1)^2)$$

- If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value, x_0, for minimum risk if the loss matrix is

$$L = \begin{bmatrix} 0 & 0.5 \\ 1.0 & 0 \end{bmatrix}$$

- The likelihood ratio is given by $l_{12} = \frac{P(x|\omega_1)}{P(x|\omega_2)}$
- The threshold is the point x_0: $l_{12} = \frac{P(x|\omega_1)}{P(x|\omega_2)} = \frac{P(\omega_2)\lambda_{21} - \lambda_{21}}{P(\omega_1)\lambda_{12} - \lambda_{11}}$
- Substituting: x_0: $exp(-x^2) = 2exp(-(x-1)^2)$
- Taking log: x_0: $-x^2 = ln2 - (x-1)^2$
- Finally: $x_0 = (1 - ln2)/2$

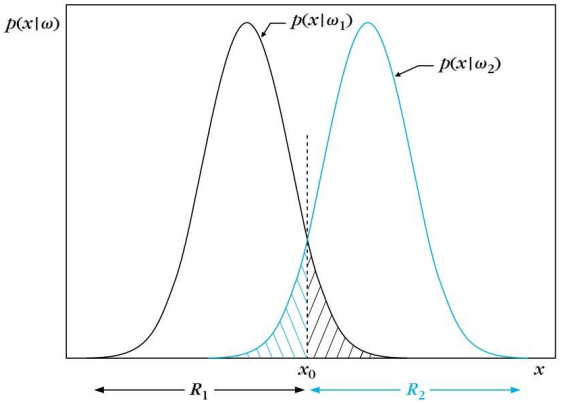
Minimising the Average Risk - Example

- The classification rule given on a previous slide as:

$$\boldsymbol{x} \in \omega_1(\omega_2) \text{ if } l_{12} = \frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)\lambda_{21} - \lambda_{22}}{P(\omega_1)\lambda_{12} - \lambda_{11}}$$

had actually become: $P(\boldsymbol{x}|\omega_1) > P(\boldsymbol{x}|\omega_2)\frac{\lambda_{21}}{\lambda_{12}}$, since $P(\omega_1) = P(\omega_2) = 1/2$ and $\lambda_{11} = \lambda_{11} = 0$.

That meant $P(\boldsymbol{x}|\omega_1)$ was multiplied by a factor less than 1 and the effect was to move the threshold to the left.



Minimising the Average Risk

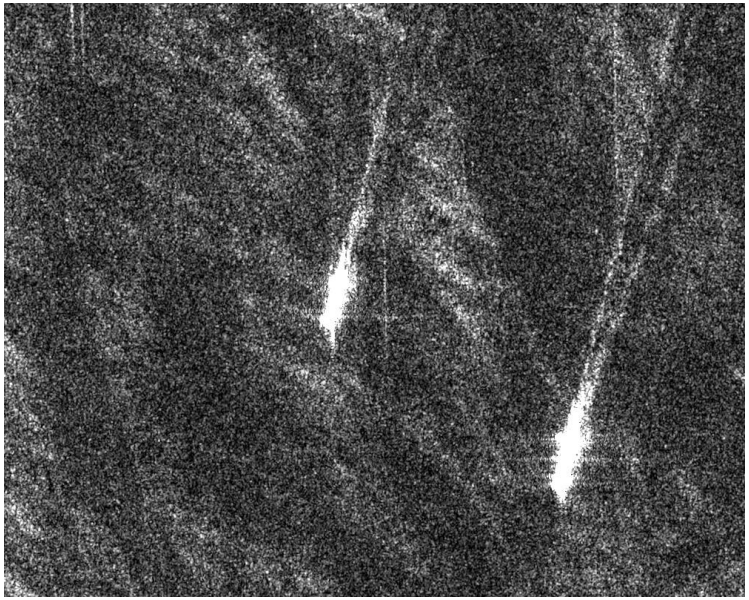
- An alternative test sometimes used is the Neyman-Pearson criterion.

$$\boldsymbol{x} \in \omega_1 \text{ if } \frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} > \theta$$

- The error for one of the classes is now constrained to be fixed and equal to a chosen value.
- Such a decision rule has been used, for example, in radar detection problems.
- The task there is to detect a target in the presence of noise.
- One type of error is the so-called false alarm—that is, to mistake the noise for a signal (target) present.
- Corresponding detectors (i.e. classifier) are called CFAR.

An Image Analysis Example: Ship detection in Synthetic Aperture Radar Images

- SAR → prime remote sensing modality for monitoring maritime activity
- Modern SAR platforms
  - Able to detect ship targets
  - Also able to detect ship wakes
- Detection of ship wakes
  - facilitates vessel detection
  - provides additional information regarding ship cruising speed & heading, size, etc



Ship targets and discernible ship wakes in NovaSAR-1 image (HH, Stripmap, source: SSTL).

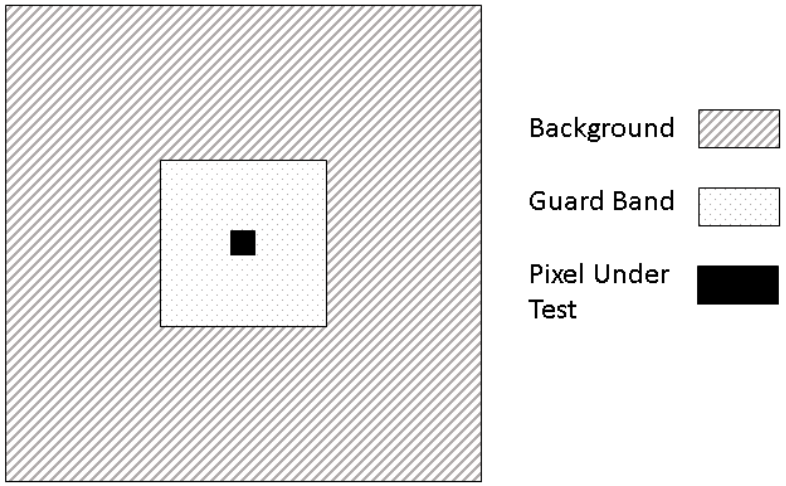
CFAR Ship Detection

Ships typically appear as bright targets in SAR images, producing high RCS (radar cross section) returns

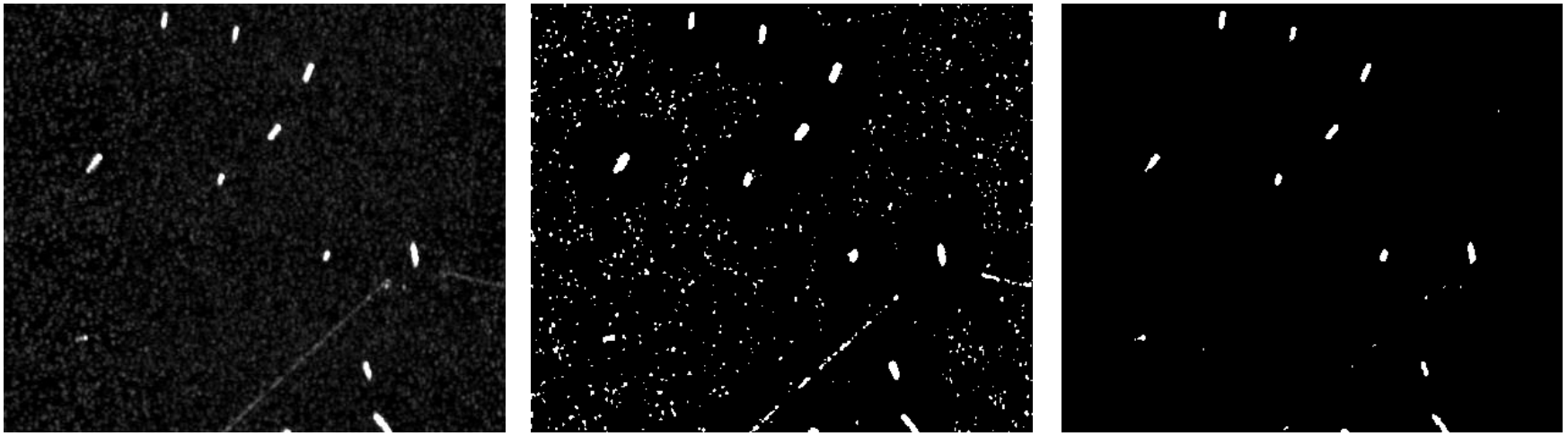
Constant False Alarm Rate (CFAR) detectors nearly ubiquitous

Improved statistical models in combination with CFAR detectors

$$P_{FA}(x) = \int_{\theta}^{\infty} p(x)dx$$



CFAR Ship detection



TerraSAR-X Data. Original image (left), Rayleigh CFAR (centre) and Rayleigh SP-CFAR (right). Detected pixels reduced by 98%.

O. Pappas, A. Achim and D. Bull, "Superpixel-Level CFAR Detectors for Ship Detection in SAR Imagery," in *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1397-1401, Sept. 2018