Non-Bayesian Methods

lecturer: Jiří Matas, `matas@cmp.felk.cvut.cz`

authors: Ondřej Drbohlav, Jiří Matas, Václav Hlaváč

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

http://cmp.felk.cvut.cz

September 2021, Ver. 1.3

Lecture Outline

1. Limitations of Bayesian Decision Theory
2. Neyman Pearson Task
3. Minimax Task
4. Wald Task

Bayesian Decision Theory

Recall:

X set of observations

K set of hidden states

D set of decisions

$p_{XK}: \; X \times K \to \mathbb{R}$: joint probability

$W: \; K \times D \to \mathbb{R}$: *loss function*,

$q: \; X \to D$: strategy

$R(q):$ risk:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \, W(k, q(x)) \tag{1}$$

Bayesian strategy q^*:

$$q^* = \operatorname*{argmin}_{q \in X \to D} R(q) \tag{2}$$

Limitations of the Bayesian Decision Theory

The limitations follow from the very ingredients of the Bayesian Decision Theory — the necessity to know all the probabilities and the loss function.

- The loss function W must make sense, but in many tasks it wouldn't
  - medical diagnosis task (W: price of medicines, staff labor, etc. but what penalty in case of patient's death?) Uncomparable penalties on different axes of X.
  - nuclear plant
  - judicial error
- The prior probabilities $p_K(k)$: must exist and be known. But in some cases it does not make sense to talk about probabilities because the events are not random.
  - $K = \{1, 2\} \equiv \{\text{own army plane}, \text{enemy plane}\}$;
    $p(x|1)$, $p(x|2)$ do exist and can be estimated, but $p(1)$ and $p(2)$ don't.
- The conditionals may be subject to non-random intervention; $p(x \mid k, z)$ where $z \in Z = \{1, 2, 3\}$ are different interventions.
  - a system for handwriting recognition: The training set has been prepared by 3 different persons. But the test set has been constructed by one of the 3 persons only. This **cannot** be done:

$$(!) \quad p(x \mid k) = \sum_z p(z) p(x \mid k, z) \tag{3}$$

Neyman Pearson Task

- $K = \{\mathsf{D}, \mathsf{N}\}$ (dangerous state, normal state)
- X set of observations
- Conditionals $p(x \mid \mathsf{D})$, $p(x \mid \mathsf{N})$ are given
- The priors $p(\mathsf{D})$ and $p(\mathsf{N})$ are unknown or do not exist
- $q \colon X \to K$ strategy

The Neyman Person Task looks for the optimal strategy q^* for which

i) the error of classification of the dangerous state is lower than a predefined threshold $\bar{\epsilon}_{\mathsf{D}}$ $(0 < \bar{\epsilon}_{\mathsf{D}} < 1)$, while

ii) the classification error for the normal state is as low as possible.

This is formulated as an optimization task with an inequality constraint:

$$q^* = \operatorname*{argmin}_{q:X\to K} \sum_{x:q(x)\neq \mathsf{N}} p(x \mid \mathsf{N}) \tag{4}$$

$$\text{subject to:} \quad \sum_{x:q(x)\neq \mathsf{D}} p(x \mid \mathsf{D}) \leq \bar{\epsilon}_{\mathsf{D}}\,. \tag{5}$$

Neyman Pearson Task

(copied from the previous slide:)

$$q^* = \underset{q:X \to K}{\operatorname{argmin}} \sum_{x:q(x) \neq \mathsf{N}} p(x \mid \mathsf{N}) \tag{4}$$

$$\text{subject to:} \quad \sum_{x:q(x) \neq \mathsf{D}} p(x \mid \mathsf{D}) \leq \bar{\epsilon}_{\mathsf{D}} \,. \tag{5}$$

A strategy is characterized by the classification error values ϵ_{N} and ϵ_{D}:

$$\epsilon_{\mathsf{N}} = \sum_{x:q(x) \neq \mathsf{N}} p(x \mid \mathsf{N}) \quad \text{(false alarm)} \tag{6}$$

$$\epsilon_{\mathsf{D}} = \sum_{x:q(x) \neq \mathsf{D}} p(x \mid \mathsf{D}) \quad \text{(overlooked danger)} \tag{7}$$

Example: Male/Female Recognition (Neyman Pearson) (1)

An aging student at CTU wants to marry. He can't afford to miss recognizing a girl when he meets her, therefore he sets the threshold on female classification error to $\bar{\epsilon}_{\mathrm{D}} = 0.2$. At the same time, he wants to minimize mis-classifying boys for girls.

- $K = \{\mathrm{D}, \mathrm{N}\} \equiv \{\mathrm{F}, \mathrm{M}\}$ (female, male)
- measurements $X = \{\text{short, normal, tall}\} \times \{\text{ultralight, light, avg, heavy}\}$
- Prior probabilities do not exist.
- Conditionals are given as follows:

$p(x|\mathrm{F})$

	u-light	light	avg	heavy
short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000

$p(x|\mathrm{M})$

	u-light	light	avg	heavy
short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169

(8)

Neyman Pearson : Solution

The optimal strategy q^* for a given $x \in X$ is constructed using the likelihood ratio $\frac{p(x \,|\, \mathsf{N})}{p(x \,|\, \mathsf{D})}$.

Let there be a constant $\mu \geq 0$. Given this μ, a strategy q is constructed as follows:

$$\frac{p(x \,|\, \mathsf{N})}{p(x \,|\, \mathsf{D})} > \mu \quad \Rightarrow \quad q(x) = \mathsf{N}\,, \tag{9}$$

$$\frac{p(x \,|\, \mathsf{N})}{p(x \,|\, \mathsf{D})} \leq \mu \quad \Rightarrow \quad q(x) = \mathsf{D}\,. \tag{10}$$

The optimal strategy q^* is obtained by selecting the minimal μ for which there still holds that $\epsilon_{\mathsf{D}} \leq \bar{\epsilon}_{\mathsf{D}}$.

Let us show this on an example.

Example: Male/Female Recognition (Neyman Pearson) (2)

$p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000

$p(x|\mathsf{M})$

	u-light	light	avg	heavy
short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169

$r(x) = p(x|\mathsf{M})/p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	0.056	0.034	0.117	0.647
normal	0.065	0.237	2.814	2.235
tall	2.000	1.750	∞	∞

rank order of $p(x|\mathsf{M})/p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	2	1	4	6
normal	3	5	10	9
tall	8	7	11	12

Here, different μ's can produce 11 different strategies.

First, let us take $2.814 < \mu < \infty$, e.g. $\mu = 3$. This produces a strategy $q^*(x) = \mathsf{F}$ everywhere except where $p(x|\mathsf{F}) = 0$. Obviously, classification error ϵ_F for F is $\epsilon_\mathsf{F} = 0$, and $\epsilon_\mathsf{M} = 1 - .255 - .169 = .576$.

Example: Male/Female Recognition (Neyman Pearson) (3)

$p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000

$p(x|\mathsf{M})$

	u-light	light	avg	heavy
short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169

$r(x) = p(x|\mathsf{M})/p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	0.056	0.034	0.117	0.647
normal	0.065	0.237	2.814	2.235
tall	2.000	1.750	∞	∞

rank, and $q^*(x) = \{\mathsf{F}, \mathsf{M}\}$ for $\mu = 2.5$

	u-light	light	avg	heavy
short	2	1	4	6
normal	3	5	10	9
tall	8	7	11	12

Next, take μ which satisfies

$$r_9 < \mu < r_{10} \quad (\text{e.g. } \mu = 2.5) \tag{11}$$

(where r_i is the likelihood ratios indexed by its rank.)

Here, $\epsilon_{\mathsf{F}} = .145$, and $\epsilon_{\mathsf{M}} = 1 - .255 - .169 - .408 = .168$.

Example: Male/Female Recognition (Neyman Pearson) (4)

$p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000

$p(x|\mathsf{M})$

	u-light	light	avg	heavy
short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169

$r(x) = p(x|\mathsf{M})/p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	0.056	0.034	0.117	0.647
normal	0.065	0.237	2.814	2.235
tall	2.000	1.750	∞	∞

rank, and $q^*(x) = \{\mathsf{F}, \mathsf{M}\}$ for $\mu = 2.1$

	u-light	light	avg	heavy
short	2	1	4	6
normal	3	5	10	9
tall	8	7	11	12

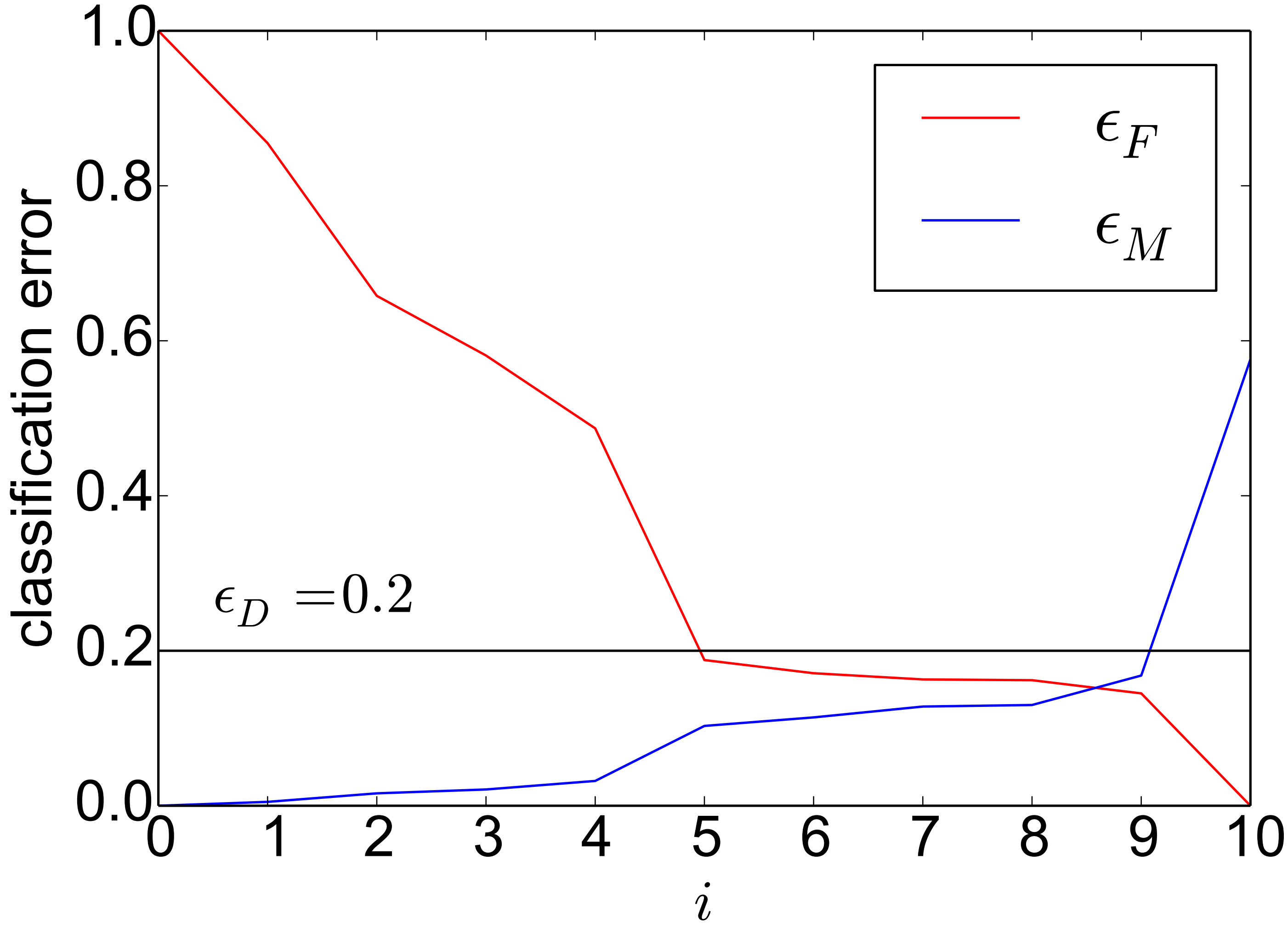
Do the same for μ satisfying

$$r_8 < \mu < r_9 \qquad (\text{e.g. } \mu = 2.1) \tag{12}$$

$\Rightarrow \epsilon_{\mathsf{F}} = .162$, and $\epsilon_{\mathsf{M}} = 0.13$.

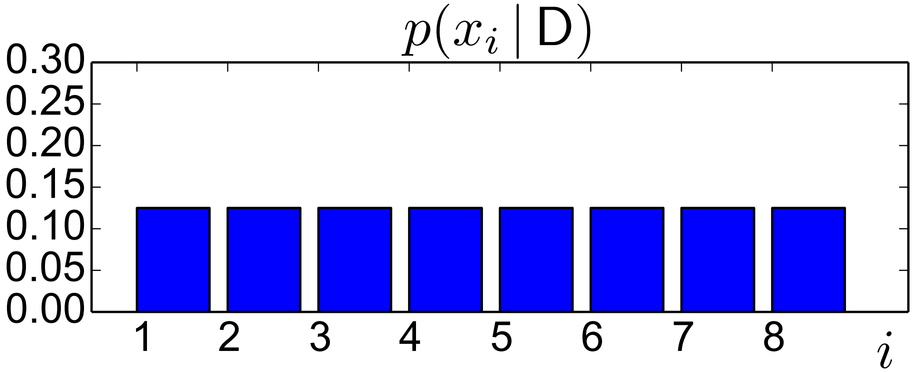
Example: Male/Female Recognition (Neyman Pearson) (5)

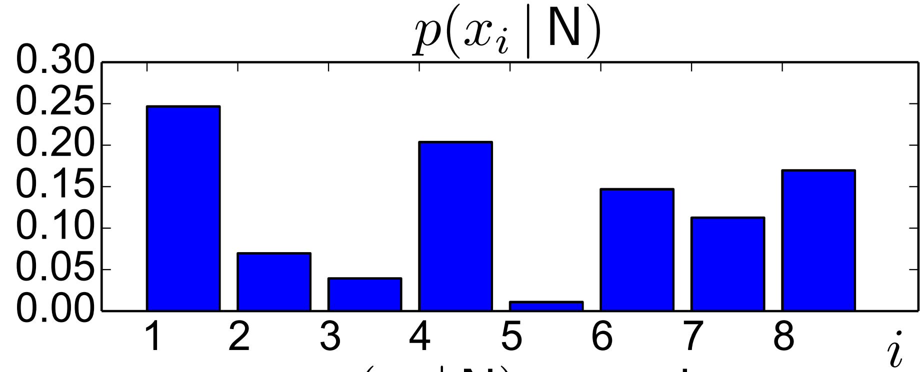
Classification errors for F and M, for $\mu_i = \frac{r_i + r_{i+1}}{2}$ and $\mu_0 = 0$.

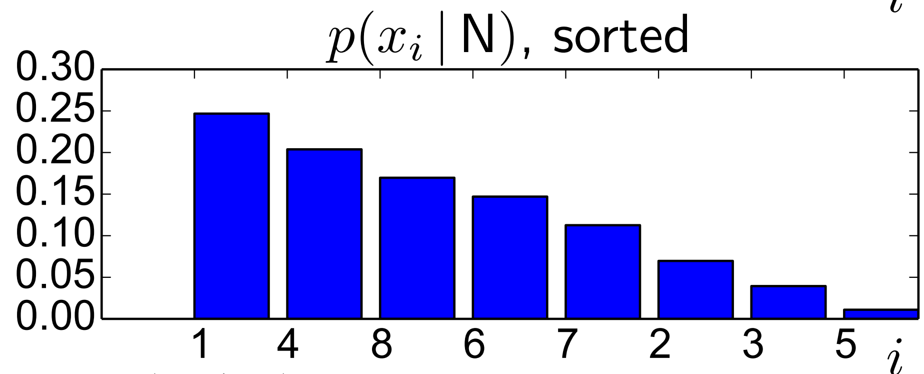


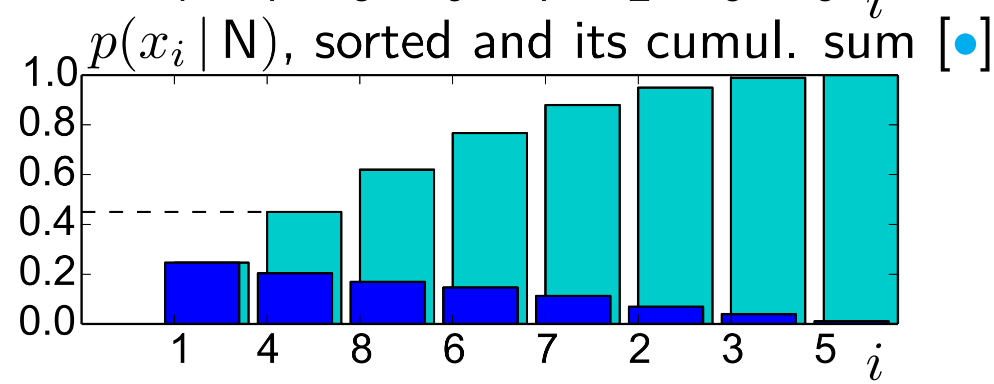
The optimum is reached for $r_5 < \mu < r_6$; $\epsilon_{\mathrm{F}} = .188$, $\epsilon_{\mathrm{M}} = .103$

Neyman Pearson : Simple Case (1)









Consider a simple case when $p(x_i \,|\, \mathsf{D}) = \text{const}$. Possible values for ϵ_{D} are $0, \frac{1}{8}, \frac{2}{8}, \ldots, 1$. If a strategy q classifies P observations as normal then $\epsilon_{\mathsf{D}} = \frac{P}{8}$.
If $P = 1$ then $\epsilon_{\mathsf{D}} = \frac{1}{8}$ and it is clear that ϵ_{N} will attain minimum if the (one) observation which is classified as normal is the one with the highest $p(x_i \,|\, \mathsf{N})$. Similarly, if $P = 2$ then the two observations to be classified as normal are the one with the first two highest $p(x_i \,|\, \mathsf{N})$. Etc.

↑ cumulative sum of sorted $p(x_i \,|\, \mathsf{N})$ shows the classification success rate for N, that is, $1 - \epsilon_{\mathsf{N}}$, for $\epsilon_{\mathsf{D}} = \frac{1}{8}, \frac{2}{8}, \ldots, 1$. For example, for $\epsilon_{\mathsf{D}} = \frac{2}{8}$ $(P = 2)$, $\epsilon_{\mathsf{N}} = 1 - 0.45 = 0.55$ (as shown, dashed.)

Neyman Pearson : Towards General Case (2)

In general, $p(x_i \,|\, \mathsf{D}) \neq$ const. Consider the following example:

$p(x_i \,|\, \mathsf{D})$

x_1	x_2	x_3
0.5	0.25	0.25

$p(x_i \,|\, \mathsf{N})$

x_1	x_2	x_3
0.6	0.35	0.05

But this can easily be converted to the previous special case by (only formally) splitting x_1 to two observations x_1' and x_1'':

$p(x_i \,|\, \mathsf{D})$

x_1'	x_1''	x_2	x_3
0.25	0.25	0.25	0.25

$p(x_i \,|\, \mathsf{N})$

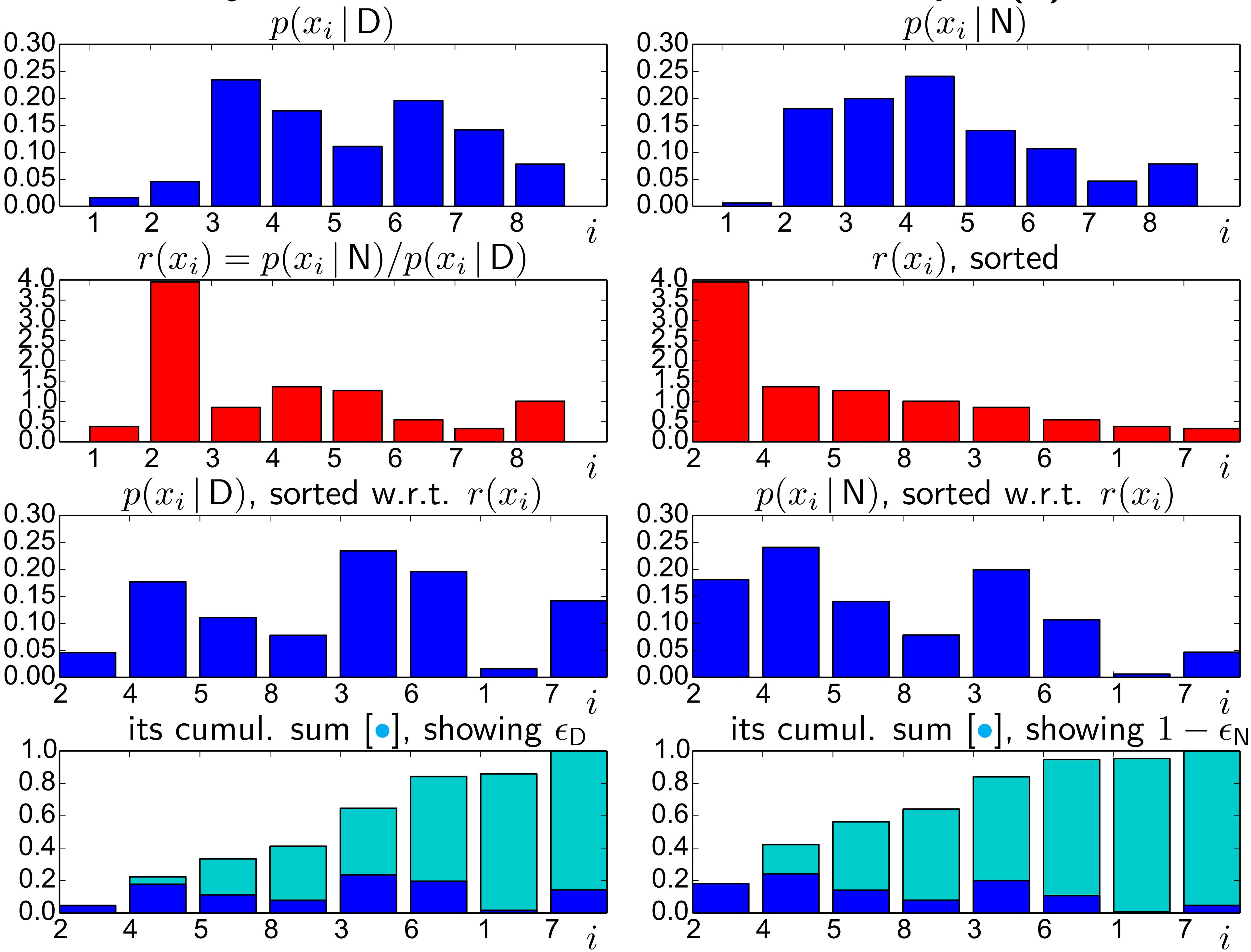
x_1'	x_1''	x_2	x_3
0.3	0.3	0.35	0.05

which would result in ordering the observations by decreasing $p(x_i \,|\, \mathsf{N})$ as: x_2, x_1, x_3.

Obviously, the same ordering is obtained when $p(x_i \,|\, \mathsf{N})$ is 'normalized' by $p(x_i \,|\, \mathsf{D})$, that is, using the likelihood ratio

$$r(x_i) = \frac{p(x_i \,|\, \mathsf{N})}{p(x_i \,|\, \mathsf{D})} \,. \tag{13}$$

Neyman Pearson : General Case Example (3)



Neyman Pearson Solution : Illustration of Principle

Lagrangian of the Neyman Pearson Task is

$$L(q) = \underbrace{\sum_{x\colon q(x)=\mathsf{D}} p(x \,|\, \mathsf{N})} + \mu \left(\sum_{x\colon q(x)=\mathsf{N}} p(x \,|\, \mathsf{D}) - \bar{\epsilon}_D \right) \tag{14}$$

$$= \overbrace{1 - \sum_{x: q(x)=\mathsf{N}} p(x \,|\, \mathsf{N})}^{=} + \mu \left(\sum_{x\colon q(x)=\mathsf{N}} p(x \,|\, \mathsf{D}) \right) - \mu\bar{\epsilon}_\mathsf{D} \tag{15}$$

$$= 1 - \mu\bar{\epsilon}_\mathsf{D} + \sum_{x\colon q(x)=\mathsf{N}} \underbrace{\{\mu\, p(x \,|\, \mathsf{D}) - p(x \,|\, \mathsf{N})\}}_{T(x)} \tag{16}$$

If $T(x)$ is negative for an x then it will decrease the objective function and the optimal strategy q^* will decide $q^*(x) = \mathsf{N}$. This illustrates why the solution to the Neyman Pearson Task has the form

$$\frac{p(x \,|\, \mathsf{N})}{p(x \,|\, \mathsf{D})} > \mu \quad \Rightarrow \quad q(x) = \mathsf{N}\,, \tag{9}$$

$$\frac{p(x \,|\, \mathsf{N})}{p(x \,|\, \mathsf{D})} \leq \mu \quad \Rightarrow \quad q(x) = \mathsf{D}\,. \tag{10}$$

Neyman Pearson : Derivation (1)

$$q^* = \min_{q:X \to K} \sum_{x:q(x) \neq \mathsf{N}} p(x \,|\, \mathsf{N}) \qquad \text{subject to:} \sum_{x:q(x) \neq \mathsf{D}} p(x \,|\, \mathsf{D}) \leq \bar{\epsilon}_{\mathsf{D}} \,. \tag{17}$$

Let us rewrite this as

$$q^* = \min_{q:X \to K} \sum_{x \in X} \alpha(x) p(x \,|\, \mathsf{N}) \qquad \text{subject to:} \quad \sum_{x \in X} [1 - \alpha(x)] p(x \,|\, \mathsf{D}) \leq \bar{\epsilon}_{\mathsf{D}} \,. \tag{18}$$

$$\text{and:} \quad \alpha(x) \in \{0, 1\} \; \forall x \in X \tag{19}$$

This is a combinatorial optimization problem. If the relaxation is done from $\alpha(x) \in \{0, 1\}$ to $0 \leq \alpha(x) \leq 1$, this can be solved by **linear programming** (LP). The Lagrangian of this problem with inequality constraints is:

$$L(\alpha(x_1), \alpha(x_2), ..., \alpha(x_N)) = \sum_{x \in X} \alpha(x) p(x \,|\, \mathsf{N}) + \mu \left(\sum_{x \in X} [1 - \alpha(x)] p(x \,|\, \mathsf{D}) - \bar{\epsilon}_{\mathsf{D}} \right) \tag{20}$$

$$- \sum_{x \in X} \mu_0(x) \alpha(x) + \sum_{x \in X} \mu_1(x) (\alpha(x) - 1) \tag{21}$$

Neyman Pearson : Derivation (2)

$$L(\alpha(x_1), \alpha(x_2), ..., \alpha(x_N)) = \sum_{x \in X} \alpha(x) p(x \,|\, \mathsf{N}) + \mu \left(\sum_{x \in X} [1 - \alpha(x)] p(x \,|\, \mathsf{D}) - \bar{\epsilon}_{\mathsf{D}} \right) \tag{20}$$

$$- \sum_{x \in X} \mu_0(x) \alpha(x) + \sum_{x \in X} \mu_1(x)(\alpha(x) - 1) \tag{21}$$

The conditions for optimality are ($\forall x \in X$):

$$\frac{\partial L}{\partial \alpha(x)} = p(x \,|\, \mathsf{N}) - \mu p(x \,|\, \mathsf{D}) - \mu_0(x) + \mu_1(x) = 0, \tag{22}$$

$$\mu \geq 0,\ \mu_0(x) \geq 0,\ \mu_1(x) \geq 0, \quad 0 \leq \alpha(x) \leq 1, \tag{23}$$

$$\mu_0(x)\alpha(x) = 0,\ \mu_1(x)(\alpha(x) - 1) = 0, \mu \left(\sum_{x \in X} [1 - \alpha(x)] p(x \,|\, \mathsf{D}) - \bar{\epsilon}_{\mathsf{D}} \right) = 0. \tag{24}$$

Case-by-case analysis:

case	implications
$\mu = 0$	L minimized by $\alpha(x) = 0 \quad \forall x$
$\mu \neq 0$, $\alpha(x) = 0$	$\mu_1(x) = 0 \Rightarrow \mu_0(x) = p(x \,\|\, \mathsf{N}) - \mu p(x \,\|\, \mathsf{D}) \Rightarrow p(x \,\|\, \mathsf{N})/p(x \,\|\, \mathsf{D}) \leq \mu$
$\mu \neq 0$, $\alpha(x) = 1$	$\mu_0(x) = 0 \Rightarrow \mu_1(x) = -[p(x \,\|\, \mathsf{N}) - \mu p(x \,\|\, \mathsf{D})] \Rightarrow p(x \,\|\, \mathsf{N})/p(x \,\|\, \mathsf{D}) \geq \mu$
$\mu \neq 0$, $0 < \alpha(x) < 1$	$\mu_0(x) = \mu_1(x) = 0 \Rightarrow p(x \,\|\, \mathsf{N})/p(x \,\|\, \mathsf{D}) = \mu$

Neyman Pearson : Derivation (3)

Case-by-case analysis:

case	implications
$\mu = 0$	L minimized by $\alpha(x) = 0 \quad \forall x$
$\mu \neq 0$, $\alpha(x) = 0$	$\mu_1(x) = 0 \Rightarrow \mu_0(x) = p(x \mid \mathsf{N}) - \mu p(x \mid \mathsf{D}) \Rightarrow p(x \mid \mathsf{N})/p(x \mid \mathsf{D}) \leq \mu$
$\mu \neq 0$, $\alpha(x) = 1$	$\mu_0(x) = 0 \Rightarrow \mu_1(x) = -[p(x \mid \mathsf{N}) - \mu p(x \mid \mathsf{D})] \Rightarrow p(x \mid \mathsf{N})/p(x \mid \mathsf{D}) \geq \mu$
$\mu \neq 0$, $0 < \alpha(x) < 1$	$\mu_0(x) = \mu_1(x) = 0 \Rightarrow p(x \mid \mathsf{N})/p(x \mid \mathsf{D}) = \mu$

Optimal Strategy for a given $\mu \geq 0$ and particular $x \in X$:

$$\frac{p(x \mid \mathsf{N})}{p(x \mid \mathsf{D})} \quad \begin{cases} < \mu & \Rightarrow q(x) = \mathsf{D} \text{ (as } \alpha(x) = 0) \\ > \mu & \Rightarrow q(x) = \mathsf{N} \text{ (as } \alpha(x) = 1) \\ = \mu & \Rightarrow \text{LP relaxation does not give the desired solution, as } \alpha \notin \{0, 1\} \end{cases} \tag{25}$$

Neyman Pearson : Note on Randomized Strategies (1)

Consider:

$p(x\|\mathsf{D})$		
x_1	x_2	x_3
0.9	0.09	0.01

$p(x\|\mathsf{N})$		
x_1	x_2	x_3
0.09	0.9	0.01

$r(x) = p(x\|\mathsf{N})/p(x\|\mathsf{D})$		
x_1	x_2	x_3
0.1	10	1

and $\bar{\epsilon}_{\mathsf{D}}$ =0.03.

- $q_1 : (x_1, x_2, x_3) \rightarrow (\mathsf{D}, \mathsf{D}, \mathsf{D}) \quad \Rightarrow \quad \epsilon_{\mathsf{D}} = 0.00,\ \epsilon_{\mathsf{N}} = 1.00$
- $q_2 : (x_1, x_2, x_3) \rightarrow (\mathsf{D}, \mathsf{D}, \mathsf{N}) \quad \Rightarrow \quad \epsilon_{\mathsf{D}} = 0.01,\ \epsilon_{\mathsf{N}} = 0.99$
- no other deterministic strategy q is feasible, that is all other ones have $\epsilon_{\mathsf{D}} > \bar{\epsilon}_{\mathsf{D}}$
- q_2 is the best deterministic strategy but it does not comply with the previous basic result of constructing the optimal strategy because it decides for N for likelihood ratio 1 but decides for D for likelihood ratios 0.01 and 10. Why is that?
- we can construct a randomized strategy which attains $\bar{\epsilon}_{\mathsf{D}}$ and reaches lower ϵ_{N}:

$$q(x_1) = q(x_3) = \mathsf{D}, \quad q(x_2) = \begin{cases} \mathsf{N} & 1/3 \text{ of the time} \\ \mathsf{D} & 2/3 \text{ of the time} \end{cases} \tag{26}$$

For such strategy, $\epsilon_{\mathsf{D}} = 0.03\,,\ \epsilon_{\mathsf{N}} = 0.7$.

Neyman Pearson : Note on Randomized Strategies (2)

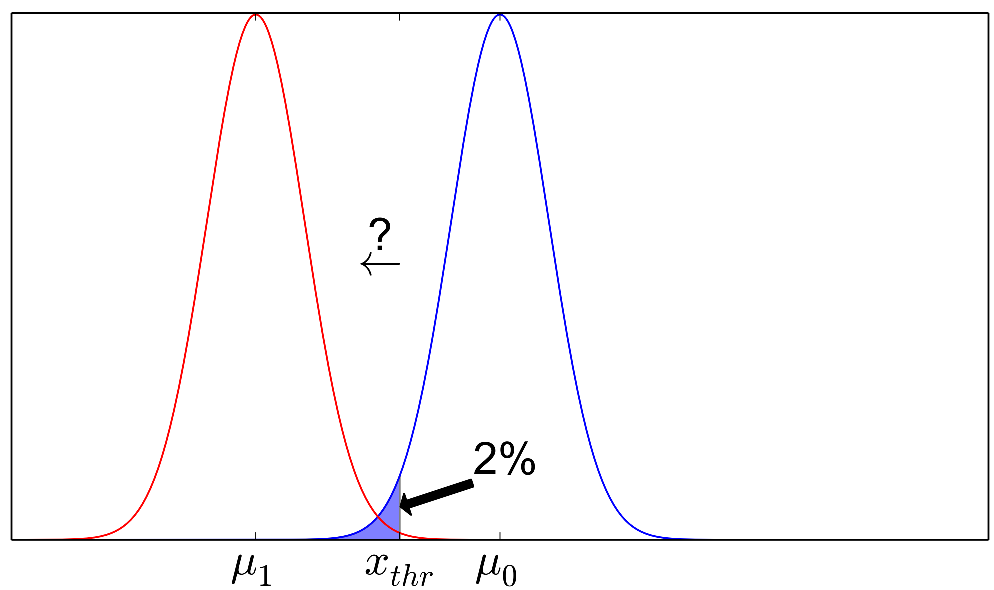
- This is not a problem but a feature which is caused by discrete nature of X (does not happen when X is continuous).
- This is exactly what the case of $\mu = p(x \,|\, \mathsf{N})/p(x \,|\, \mathsf{D})$ is on slide 18.

Neyman Pearson : Notes (1)

- The task can be generalized to 3 hidden states, of which 2 are dangerous, $K = \{\mathrm{N}, \mathrm{D}_1, \mathrm{D}_2\}$. It is formulated as an analogous problem with two inequality constraints and minimization of classification error for N.
- Neyman's and Pearson's work dates to 1928 and 1933.
- A particular strength of the approach lies in that the likelihood ratio $r(x)$ or even $p(x \,|\, \mathrm{N})$ need not be known. For the task to be solved, it is enough to know the $p(x \,|\, \mathrm{D})$ and the **rank order** of the likelihood ratio
(to be demonstrated on the next page)

Neyman Pearson : Notes (2)

- Consider a medicine for reducing weight. The normal population has a distribution of weight $p(x \mid \mathsf{D})$ as shown in blue. Let it be normal, $p(x \mid \mathsf{D}) = \mathcal{N}(x \mid \mu_0, \sigma)$. The distribution of weights after 1 month of taking the medicine is assumed to be normal as well, with the same variance but uknown shift of mean to the left, $p(x \mid \mathsf{N}) = \mathcal{N}(x \mid \mu_1, \sigma)$, with $\mu_1 < \mu_0$ but otherwise unknown (shown in red).
  The likelihood ratio is
  $r(x) = \exp \frac{1}{2\sigma^2}\left(-(x-\mu_1)^2 + (x-\mu_0)^2\right) = \exp\left(\frac{1}{\sigma^2}(\mu_1 - \mu_0)x + \text{const}\right)$. It is thus decreasing (monotone) with x (irrespective of μ_1, $\mu_1 < \mu_0$).
- Setting $\bar{\epsilon}_{\mathsf{D}} = 0.02$, we go along the decreasing $r(x)$ and find the point x_{thr} for which $\int_{-\infty}^{x_{thr}} p(x \mid \mathsf{D}) = \bar{\epsilon}_{\mathsf{D}} = 0.02$ (0.02-quantile). Note that the threshold μ on $r(x)$ is still uknown as $p(x \mid \mathsf{N})$ is unknown.



Minimax Task

- $K = \{1, 2, .., N\}$
- X set of observations
- Conditionals $p(x \mid k)$ are known $\forall k \in K$
- The priors $p(k)$ are unknown or do not exist
- $q \colon X \to K$ strategy

The Minimax Task looks for the optimum strategy q^* which minimizes the classification error of the worst classified class:

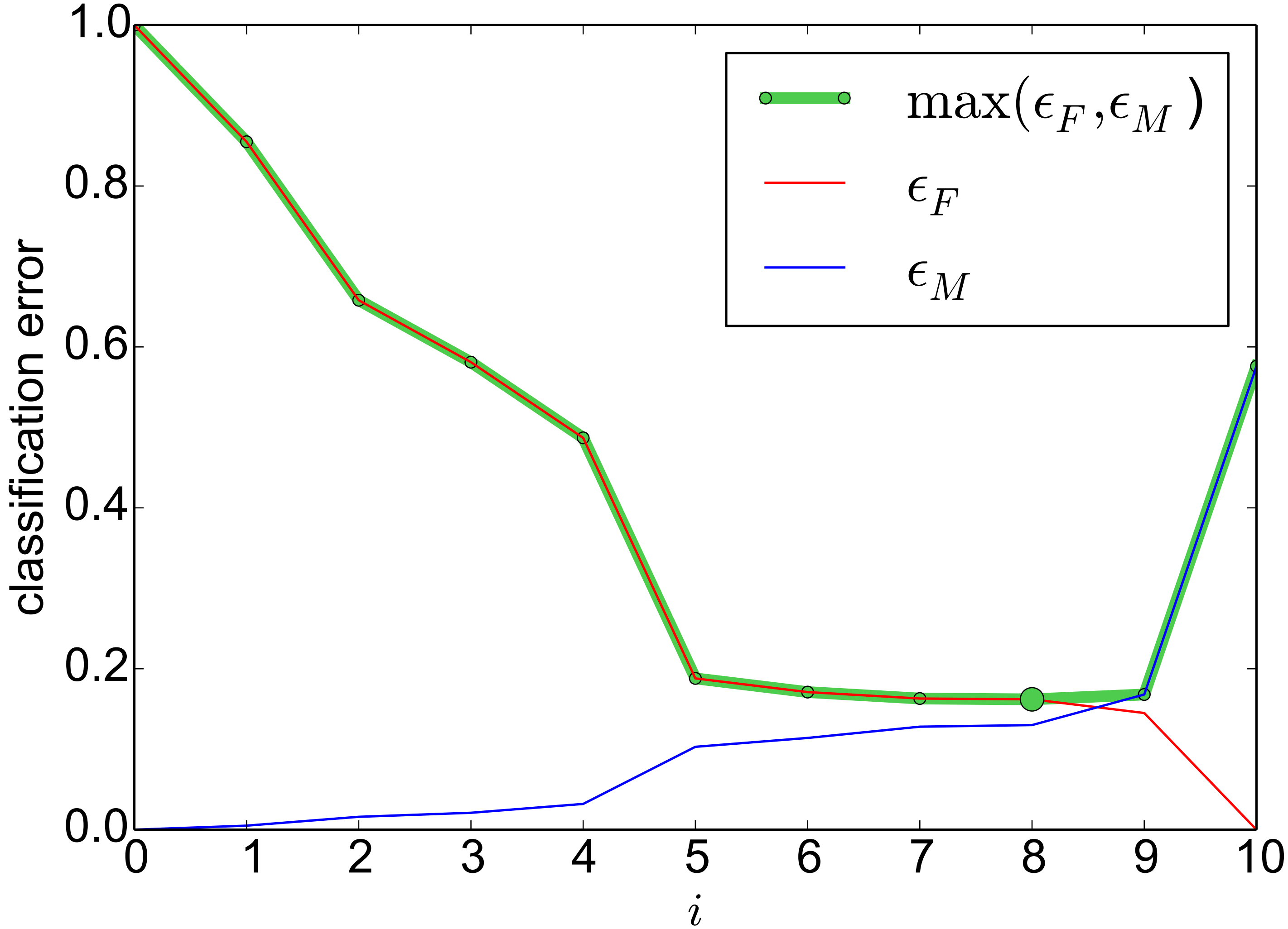
$$q^* = \operatorname*{argmin}_{q:X \to K} \max_{k \in K} \epsilon(k), \quad \text{where} \tag{27}$$

$$\epsilon(k) = \sum_{x \colon q(x) \neq k} p(x \mid k) \tag{28}$$

- Example: A recognition algorithm qualifies for a competition using preliminary tests. During the final competition, only objects from the hardest-to-classify class are used.
- For a 2-class problem, the strategy is again constructed using the likelihood ratio.
- In the case of continuous observations space X, equality of classification errors is attained: $\epsilon_1 = \epsilon_2$
- The derivation can again be done using Linear Programming.

Example: Male/Female Recognition (Minimax)

Classification errors for F and M, for $\mu_i = \frac{r_i + r_{i+1}}{2}$ and $\mu_0 = 0$.



The optimum is attained for $i = 8$, $\epsilon_{\mathsf{F}} = .162$, $\epsilon_{\mathsf{M}} = .13$. The corresponding strategy is as shown on slide 11.

Minimax: Comparison with Bayesian Decision with Unknown Priors

- Consider the same setting as in the Minimax task, but let the priors $p(k)$ exist but be unknown.
- The Bayesian error ϵ for strategy q is

$$\epsilon = \sum_k \sum_{x\colon q(x) \neq k} p(x, k) = \sum_k p(k) \underbrace{\sum_{x\colon q(x) \neq k} p(x \mid k)}_{\epsilon(k)} \tag{29}$$

- We want to minimize ϵ but we do not know $p(k)$'s. What is the maximum it can attain? Obviously, the $p(k)$'s do the convex combination of the class errors $\epsilon(k)$; the maximum Bayesian error will be attained when $p(k) = 1$ for the class k with the highest class error $\epsilon(k)$.
- Thus, to minimize the Bayesian error ϵ under this setting, the solution is to minimize the error of the hardest-to-classify class.
- Therefore, Minimax formulation and the Bayesian formulation with Unknown Priors lead to the same solution.

Wald Task (1)

- Let us consider classification with two states, $K = \{1, 2\}$.
- We want to set a threshold ϵ on the classification error of both of the classes: $\epsilon_1 \leq \epsilon$, $\epsilon_2 \leq \epsilon$.
- It is clear that there may be **no** feasible solution if ϵ is set too low.
- That is why the possibility of decision "do not know" is introduced. Thus $D = K \cup \{?\}$
- A strategy $q : X \rightarrow D$ is characterized by:

$$\epsilon_1 = \sum_{x:\, q(x)=2} p(x \,|\, 1) \quad \text{(classification error for 1)} \tag{30}$$

$$\epsilon_2 = \sum_{x:\, q(x)=1} p(x \,|\, 2) \quad \text{(classification error for 2)} \tag{31}$$

$$\kappa_1 = \sum_{x:\, q(x)=?} p(x \,|\, 1) \quad \text{(undecided rate for 1)} \tag{32}$$

$$\kappa_2 = \sum_{x:\, q(x)=?} p(x \,|\, 2) \quad \text{(undecided rate for 2)} \tag{33}$$

Wald Task (2)

- The optimal strategy q^*:

$$q^* = \operatorname*{argmin}_{q:X\to D} \max_{i=\{1,2\}} \kappa_i \tag{34}$$

$$\text{subject to: } \epsilon_1 \leq \epsilon,\ \epsilon_2 \leq \epsilon \tag{35}$$

- The task is again solvable using LP (even for more than 2 classes)
- The optimal solution is again based on the likelihood ratio

$$r(x) = \frac{p(x\,|\,1)}{p(x\,|\,2)} \tag{36}$$

- The optimal strategy is constructed using suitably chosen thresholds μ_l and μ_h such that:

$$q(x) = \begin{cases} 2 & \text{for } r(x) < \mu_l \\ 1 & \text{for } r(x) > \mu_h \\ ? & \text{for } \mu_l \leq r(x) \leq \mu_h \end{cases} \tag{37}$$

Example: Male/Female Recognition (Wald)

Solve the Wald task for $\epsilon = 0.05$.

$p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000

$p(x|\mathsf{M})$

	u-light	light	avg	heavy
short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169

$r(x) = p(x|\mathsf{M})/p(x|\mathsf{F})$

	u-light	light	avg	heavy
short	0.056	0.034	0.117	0.647
normal	0.065	0.237	2.814	2.235
tall	2.000	1.750	∞	∞

rank, and $q^*(x) = \{\mathsf{F}, \mathsf{M}, ?\}$

	u-light	light	avg	heavy
short	2	1	4	6
normal	3	5	10	9
tall	8	7	11	12

Result: $\epsilon_\mathsf{M} = 0.032$, $\epsilon_\mathsf{F} = 0$, $\kappa_\mathsf{M} = 0.544$, $\kappa_\mathsf{F} = 0.487$

$(r_4 < \mu_l < r_5,\ r_{10} < \mu_h < \infty)$