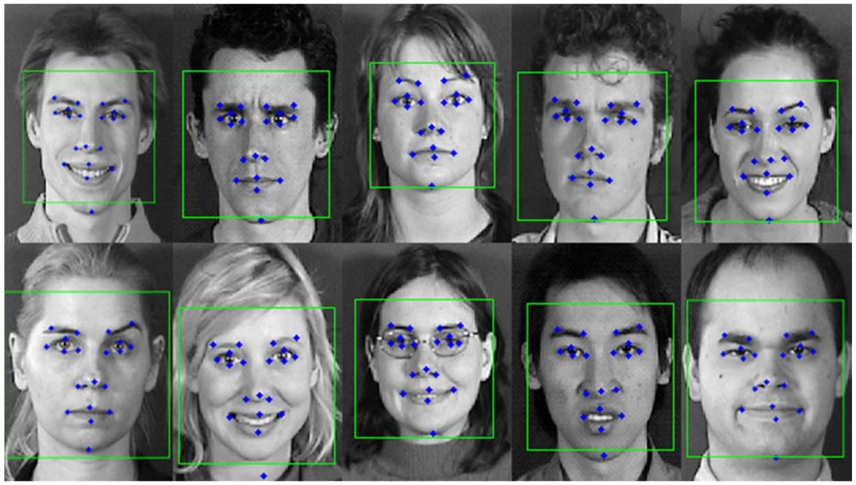
Learning from Data
5. Classification: Performance Metrics

Spring 2017

Classification

- Supervised (preditive) learning: given examples with labels, predict labels for all unseen examples
  - Classification:
    - label = category,
    - $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, m\}$, m = number of classes
    - $\ell(\mathbf{x}, y, h) = 1(h(\mathbf{x}) \neq y)$, Risk = $P(Y \neq h(\mathbf{X})) = P(\text{Error})$



$\mathbf{x}$ = facial geometry features
y = gender label

Confusion Matrix or Contingency Table

		Truth				
		$y=1$	$\ldots$	$y=j$	$\ldots$	$y=m$
Decision	$\widehat{y}=h(\mathbf{x})=1$	$\widehat{n}_{11}$	$\ldots$	$\widehat{n}_{1j}$	$\ldots$	$\widehat{n}_{1m}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\widehat{y}=h(\mathbf{x})=i$	$\widehat{n}_{i1}$	$\ldots$	$\widehat{n}_{ij}$	$\ldots$	$\widehat{n}_{im}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\widehat{y}=h(\mathbf{x})=m$	$\widehat{n}_{m1}$	$\ldots$	$\widehat{n}_{mj}$	$\ldots$	$\widehat{n}_{mm}$

$\widehat{n}_{ij}$ = count of the total number of class j samples which are classified by $h(\mathbf{x})$ as class i.

- $n =$ total number of samples $= \sum_{i=1}^{m} \sum_{j=1}^{m} \hat{n}_{ij}$
- Correct Classification Rate (CCR) $= \sum_{j=1}^{m} \hat{n}_{jj}/n$

Error Rates for Binary Classification ($m = 2$)

- Class 0 = Negative or Null Hypothesis
- Class 1 = Positive or Alternative Hypothesis
- $n =$ total number of samples
- $n_+ =$ true total number of positives
- $n_- =$ true total number of negatives
- $\hat{n}_+ =$ decided total number of positives
- $\hat{n}_- =$ decided total number of negatives
- T = True, F = False, P = Positive, N = Negative, R = Rate
- Error Rates = normalized error counts = empirical estimates of conditional error probabilities
- TP = count of True Positives, TN, FP, FN similar

Error Rates for Binary Classification ($m = 2$)

- Prevalence = n_+/n
- TPR = True Positive Rate = Sensitivity = Recall = Hit Rate = Detection Rate = "Power" of decision rule = TP/n_+ = estimate of: $P(h(\mathbf{x}) = 1|Y = 1)$
- FPR = False Positive Rate = False Alarm (FA) Rate = Type I Error Rate = "Size" of decision rule = FP/n_- = estimate of: $P(h(\mathbf{x}) = 1|Y = 0)$
  - Detection, False Alarm: used in Communications, Radar
  - When prevalence is low (rare event), FPR will be very small. Then FP is more meaningful than FPR
- Positive Likelihood Ratio (LR+): TPR/FPR
- Negative Likelihood Ratio (LR-): FNR/TNR
- Diagnostic Odds Ratio (DOR): LR+/LR-

Error Rates for Binary Classification ($m = 2$)

		Truth			
		$y = 1$	$y = 0$	Row sums	
Decision	$\widehat{y} = h(\mathbf{x}) = 1$	TP	FP	$\widehat{n}_+ = \mathrm{TP} + \mathrm{FP}$	Decided total numbers
	$\widehat{y} = h(\mathbf{x}) = 0$	FN	TN	$\widehat{n}_- = \mathrm{FN} + \mathrm{TN}$	
	Column sums:	$n_+ = \mathrm{TP} + \mathrm{FN}$	$n_- = \mathrm{FP} + \mathrm{TN}$	$n = \mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}$	
		True total numbers			

		Truth	
		$y = 1$	$y = 0$
Decision	$\widehat{y} = h(\mathbf{x}) = 1$	$\mathrm{TP}/n_+ = \mathrm{TPR} = \text{sensitivity} = \text{recall}$	$\mathrm{FP}/n_- = \mathrm{FPR} = \text{type I}$
	$\widehat{y} = h(\mathbf{x}) = 0$	$\mathrm{FN}/n_+ = \mathrm{FNR} = \text{miss rate} = \text{type II}$	$\mathrm{TN}/n_- = \mathrm{TNR} = \text{specificity}$

Error Rates for Binary Classification ($m = 2$)

		Truth	
		$y = 1$	$y = 0$
Decision	$\widehat{y} = h(\mathbf{x}) = 1$	$\mathrm{TP}/\widehat{n}_+ = \text{precision} = \mathrm{PPV}$	$\mathrm{FP}/\widehat{n}_+ = \mathrm{FDR}$
	$\widehat{y} = h(\mathbf{x}) = 0$	$\mathrm{FN}/\widehat{n}_- = \mathrm{FOR}$	$\mathrm{TN}/\widehat{n}_- = \mathrm{NPV}$

PPV = Positive Predictive Value, FDR = False Discovery Rate, FOR = False Omission Rate, NPV = Negative Predictive Value

- Precision = $\mathrm{TP}/\hat{n}_+$= estimate of: $P(Y = 1|h(\mathbf{x}) = 1)$
  - focuses on positives
  - useful when notion of negative unclear
  - used in information retrieval systems (used in conjunction with recall)

Error Rates for Binary Classification ($m = 2$)

- F-score or F_1-score combines precision (P) and recall (R) into a single statistic via their harmonic mean:

  $$\mathrm{F}_1{}^{-1} = \tfrac{1}{2}(\mathrm{P}^{-1}+\mathrm{R}^{-1}) \text{ or } \mathrm{F}_1 = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$$

  – widely used in information retrieval systems

  – Why harmonic mean instead of arithmetic mean? Consider following example:

    - $\mathrm{P} = 10^{-4}, \mathrm{R} \approx 1 \Rightarrow \frac{\mathrm{P}+\mathrm{R}}{2} \approx 0.5$, but $\mathrm{F}_1 = \frac{2\times 10^{-4}\times 1}{1+10^{-4}} \approx 0.002$

Generalization of rates to multiple classes

- macro-averaging:

$$\frac{1}{m}\sum_{j=1}^{m} \mathrm{Rate}(j),$$

where $\mathrm{Rate}(j)$ = error rate from class j's binary contingency table where class j is positive and all other classes together are negative

- micro-averaging: pool together counts from the binary contingency tables of all classes and then compute rate

Generalization of rates to multiple classes

Class 1

	$y = 1$	$y \neq 1$
$\widehat{y} = 1$	TP_1	FP_1
$\widehat{y} \neq 1$	FN_1	TN_1

$\ldots$

Class m

	$y = m$	$y \neq m$
$\widehat{y} = m$	TP_m	FP_m
$\widehat{y} \neq m$	FN_m	TN_m

Pooled

	y	not y
$\widehat{y}$	$\sum_{j=1}^{m} \mathrm{TP}_j$	$\sum_{j=1}^{m} \mathrm{FP}_j$
not $\widehat{y}$	$\sum_{j=1}^{m} \mathrm{FN}_j$	$\sum_{j=1}^{m} \mathrm{TN}_j$

Illustration of difference between macro- and micro- averaging (for precision).
Macro-averaged precision $= \frac{1}{m} \sum_{j=1}^{m} \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}$.
Micro-averaged precision $= \frac{\sum_{j=1}^{M} \mathrm{TP}_j}{\sum_{j=1}^{m} (\mathrm{TP}_j + \mathrm{FP}_j)}$.

Confidence Intervals

- Error Rates = normalized error counts = empirical estimates of conditional error probabilities.
- It is considered good practice to report the estimate of an error rate together with a 1, 2, or 3 sigma confidence interval
- Example: $\mathrm{TPR} = \mathrm{TP}/n_+$ = estimate of: $P_D = P(h(\mathbf{x}) = 1 | Y = 1)$
  - Now, $\mathrm{TPR} = \frac{1}{n_+}\sum_{j:y_j=1} \hat{Y}_j$ is a random variable with mean P_D and variance $\frac{1}{n_+}P_D(1 - P_D)$ since $\hat{Y}_j$'s are IID Bernoulli random variables with mean P_D.
  - An estimate of P_D is given by TPR
  - An estimate of the standard deviation of TPR is given by $\hat{\sigma}_{\mathrm{TPR}} = \sqrt{\frac{\mathrm{TPR}(1-\mathrm{TPR})}{n_+}}$
  - Thus we report the estimate of P_D as: $\mathrm{TPR} \pm k\hat{\sigma}_{\mathrm{TPR}}$, where k = 1 for 68% confidence, k = 2 for 95% confidence and k = 3 for 99% confidence

Receiver Operating Characteristic (ROC)

- Associated with a decision rule $h(\mathbf{x})$ are its
  - Detection probability: $P_D(h) = P(h(\mathbf{X}) = 1|Y = 1)$ and
  - False alarm probability: $P_{FA}(h) = P(h(\mathbf{X}) = 1|Y = 0)$

- The overall error probability can be expressed in terms of these two numbers:

$$\begin{aligned} P_{\text{error}} &= P(h(\mathbf{X}) \neq Y) \\ &= P(Y = 0)P_{FA}(h) + P(Y = 1)(1 - P_D(h)) \end{aligned}$$

- Associated with a family of decision rules: $\mathcal{H} = \{h\}$ is the set $\{(P_{FA}(h), P_D(h)) : h \in \mathcal{H}\}$ of pairs of detection and false-alarm probabilities of these decision rules

Receiver Operating Characteristic (ROC)

- Example:

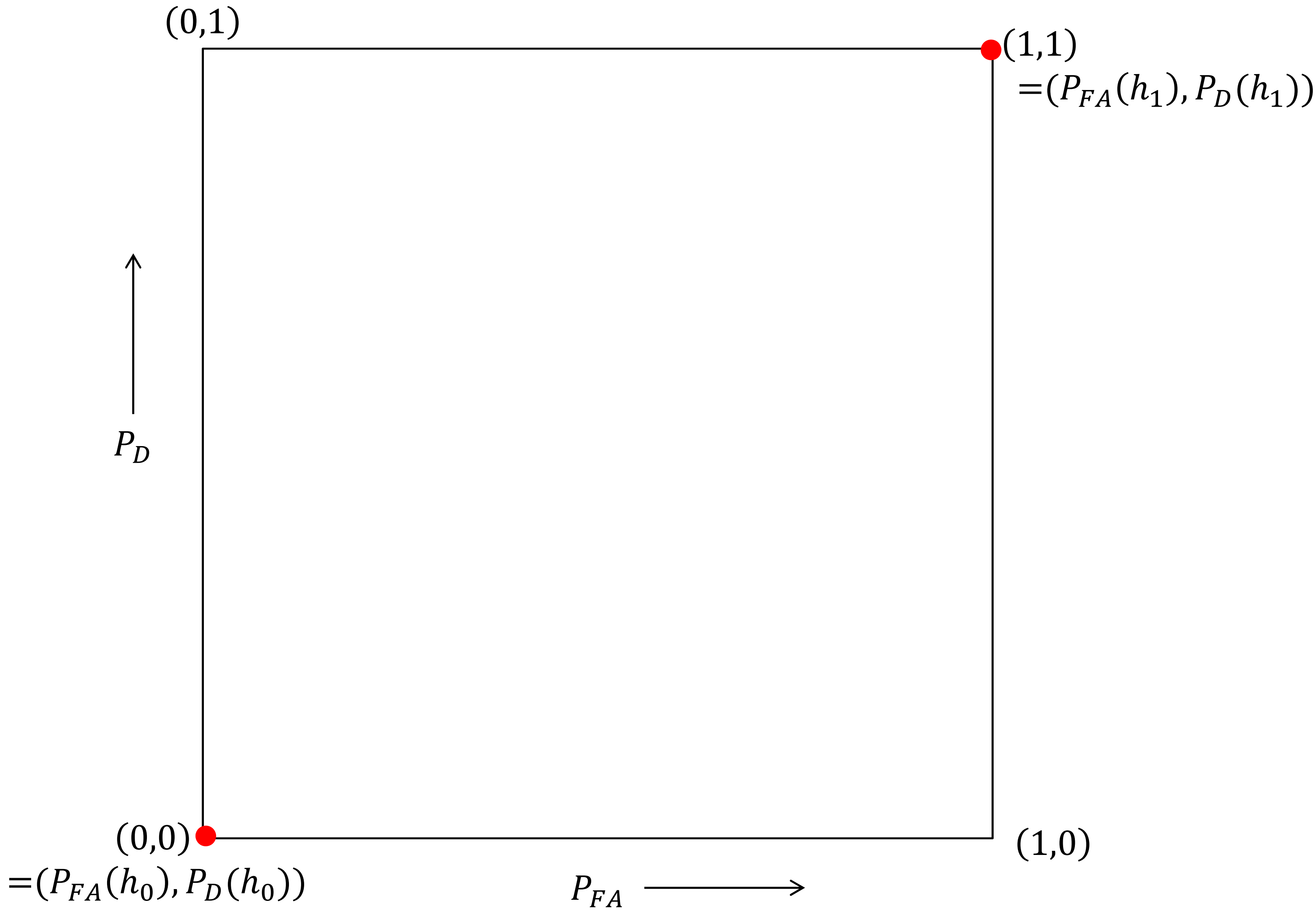
$$\mathcal{H}_{\text{trivial}} = \{h_0(\mathbf{x}) \equiv 0, h_1(\mathbf{x}) \equiv 1\}$$

the family of trivial decision rules:

  - Always decide zero irrespective of the value of $\mathbf{x}$: $h_0(\mathbf{x}) = 0\ \forall \mathbf{x}, P_{FA}(h_0) = P_D(h_0) = 0.$

  - Always decide one irrespective of the value of $\mathbf{x}$: $h_1(\mathbf{x}) = 1\ \forall \mathbf{x}, P_{FA}(h_1) = P_D(h_1) = 1.$

Receiver Operating Characteristic (ROC)



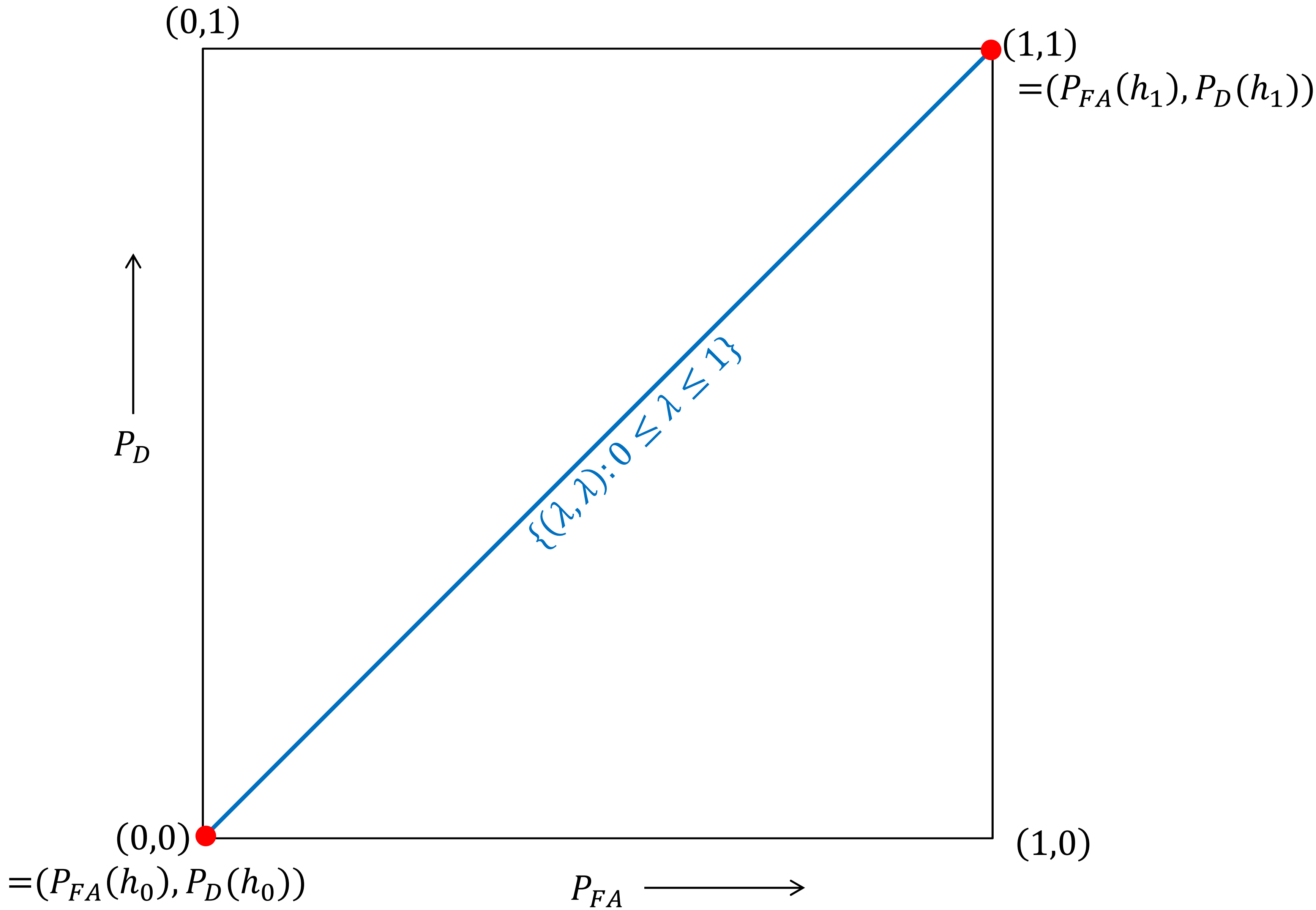
Receiver Operating Characteristic (ROC)

- Randomized decision rules:
  - Given: a family of decision rules $\mathcal{H} = \{h\}$
  - Randomized decision rule: randomly select a rule H from $\mathcal{H}$ according to some distribution $p(h)$

- Detection probability = $E_H[P_D(H)], H \sim p(h)$
- False alarm probability = $E_H[P_{FA}(H)], H \sim p(h)$

- Example: For $\mathcal{H}_{\text{trivial}} = \{h_0(\mathbf{x}) \equiv 0, h_1(\mathbf{x}) \equiv 1\}$,
  - The set of all randomized decision rules of this family can be described as $h_Z(\mathbf{x})$, where $P(Z = 1) = \lambda, P(Z = 0) = 1 - \lambda$, and $\lambda \in [0,1]$.
  - $P_{FA}(h_Z) = \lambda P_{FA}(h_1) + (1 - \lambda)P_{FA}(h_0) = \lambda.$
  - $P_D(h_Z) = \lambda P_D(h_1) + (1 - \lambda)P_D(h_0) = \lambda.$
  - As λ ranges from 0 to 1, the pair $(P_{FA}, P_D) = (\lambda, \lambda)$ traces out a straight line from $(0,0)$ and to $(1,1)$

Receiver Operating Characteristic (ROC)



Receiver Operating Characteristic (ROC)

- Convex hull of $\{(P_{FA}(h), P_D(h)) : h \in \mathcal{H} \cup \mathcal{H}_{\text{trivial}}\}$: is the set of (P_{FA}, P_D) pairs of all randomized decision rules of the family (including the trivial decision rules)

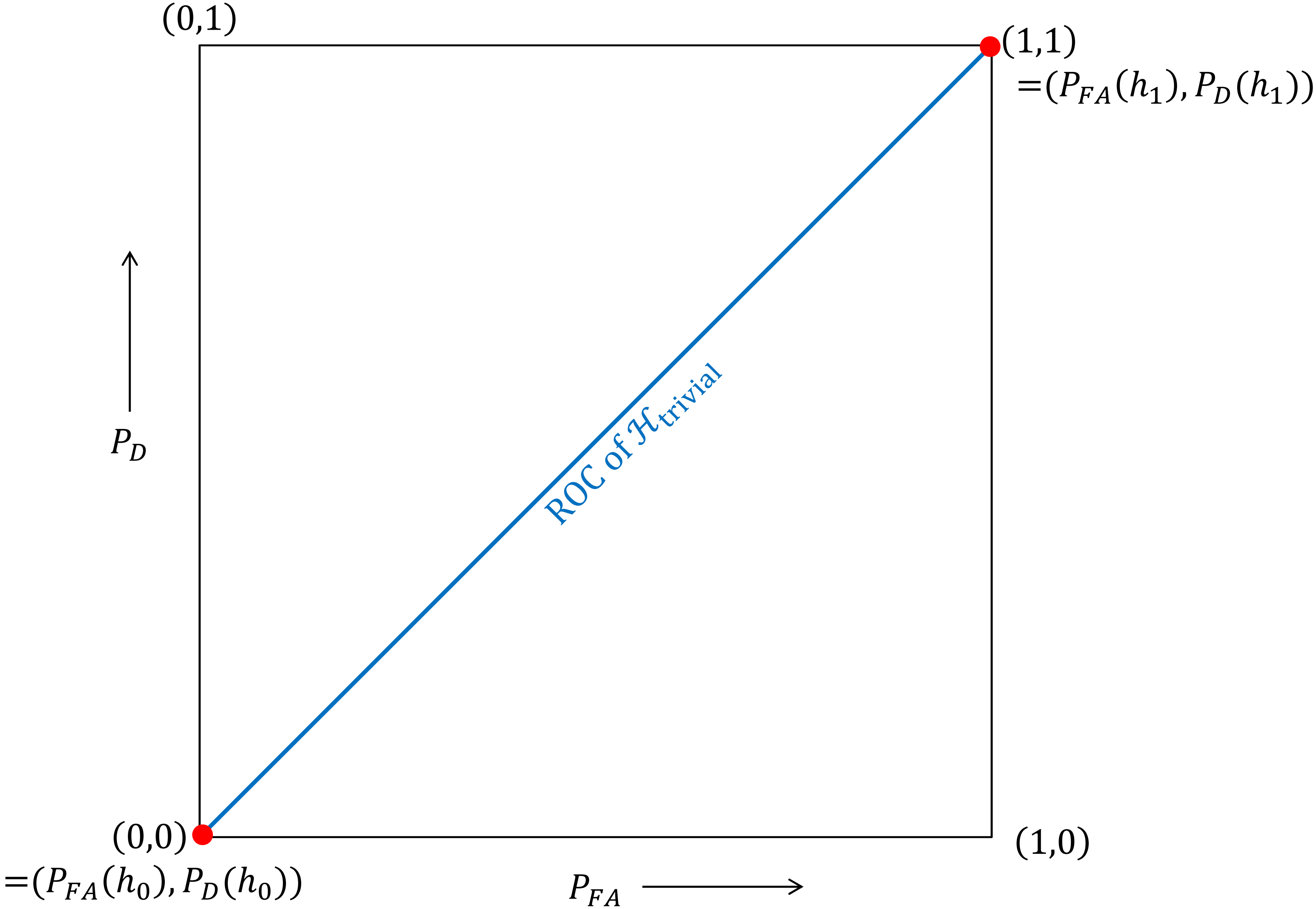
$$\text{conv}(\{(P_{FA}(h), P_D(h)) : h \in \mathcal{H} \cup \mathcal{H}_{\text{trivial}}\})$$

$$= \{(E[P_{FA}(H)], E[P_D(H)]) : H \text{ a RV over} \mathcal{H} \cup \mathcal{H}_{\text{trivial}}\}$$

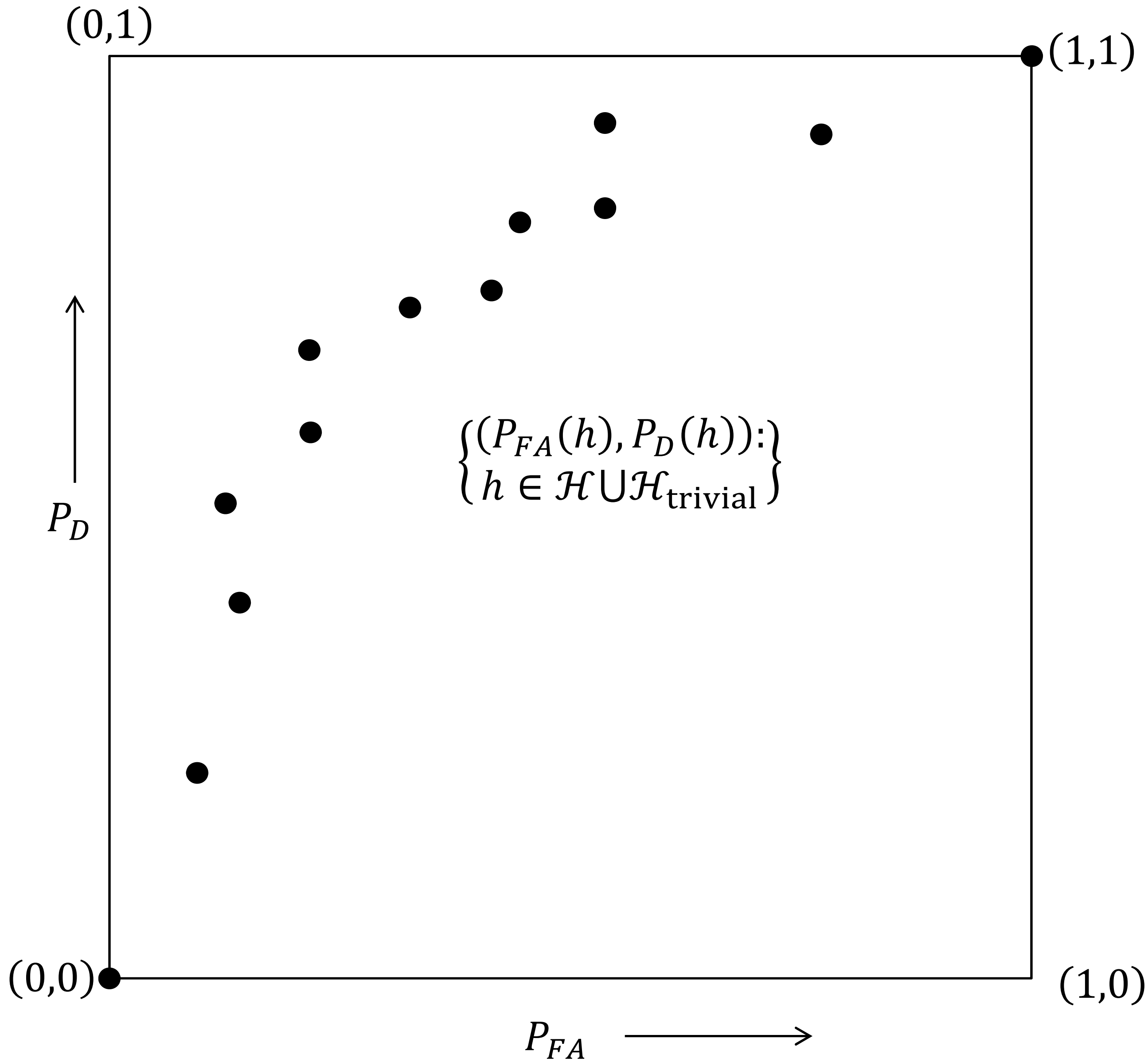
- i.,e., the set of (P_{FA}, P_D)pairs obtained by taking all possible averages of (P_{FA}, P_D) pairs of the rules in the family (including the trivial decision rules)

- ROC curve (terminology from Radar): is the upper envelope of the convex hull

Receiver Operating Characteristic (ROC)

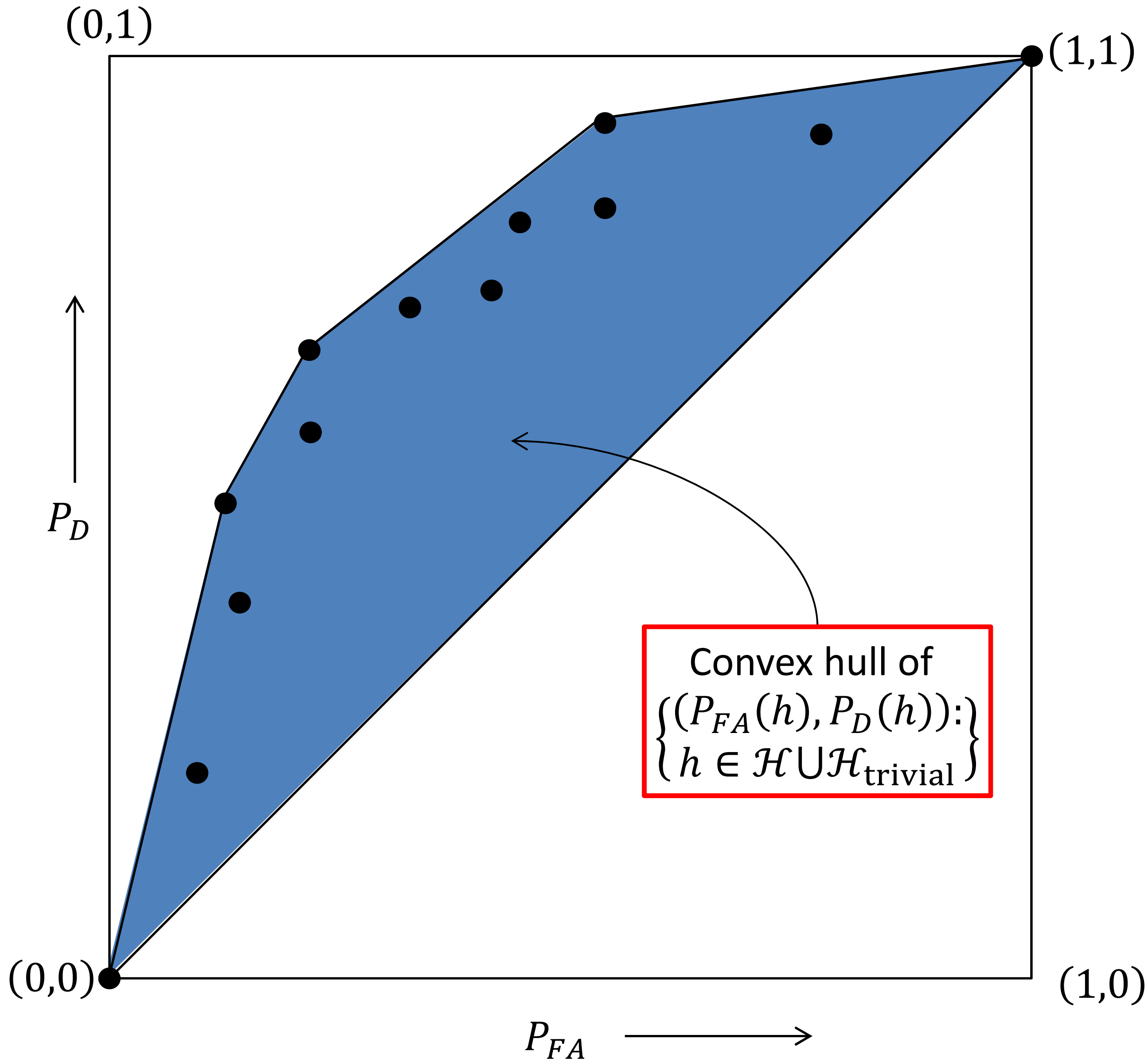


ROC



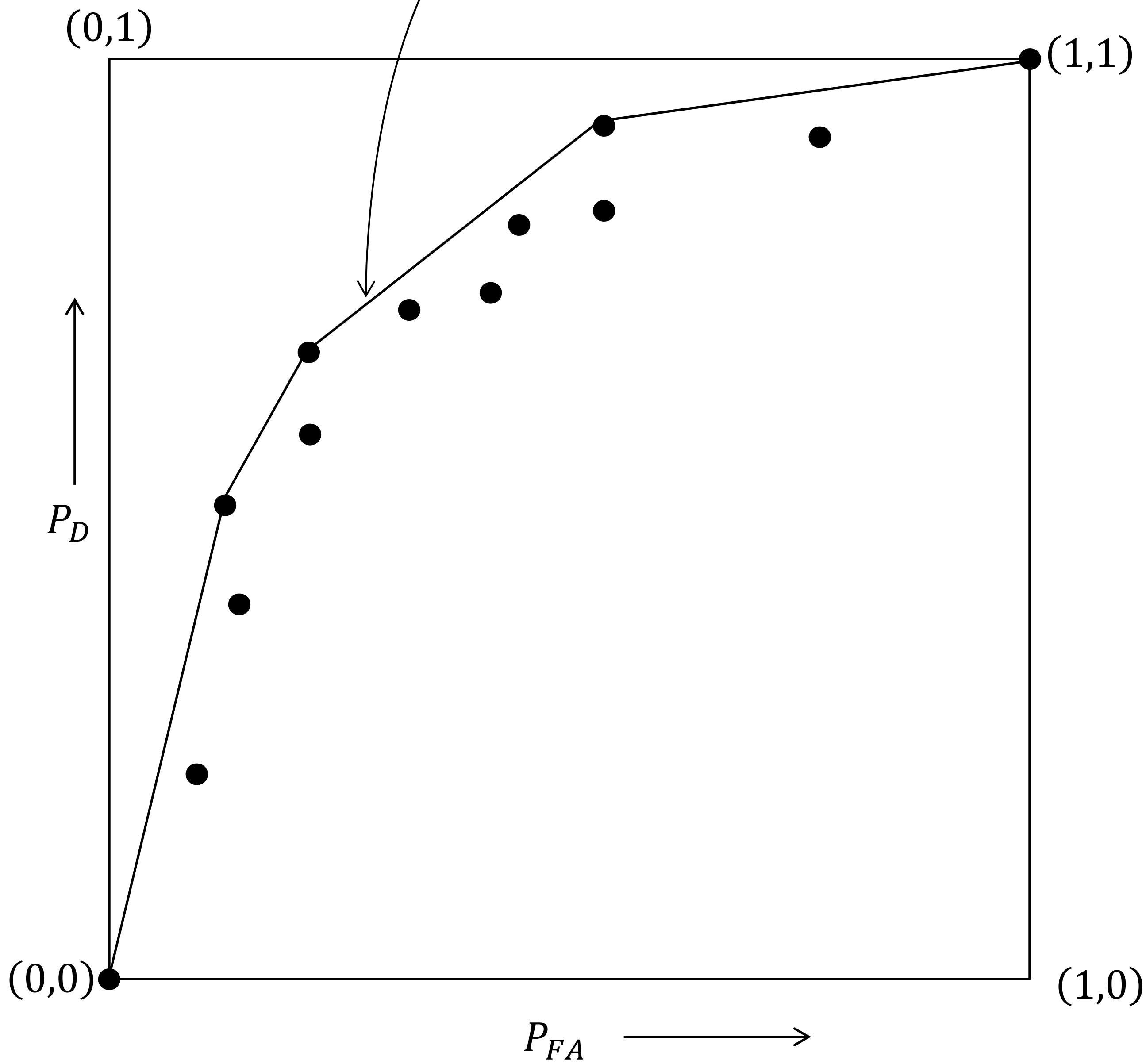
(0,1)
(1,1)
$\left\{ \begin{array}{l} (P_{FA}(h), P_D(h)): \\ h \in \mathcal{H} \cup \mathcal{H}_{\text{trivial}} \end{array} \right\}$
P_D
(0,0)
(1,0)
P_{FA}

ROC


(0,1)
(1,1)
(0,0)
(1,0)
P_D
P_{FA}
Convex hull of
$\left\{ \begin{array}{c} (P_{FA}(h), P_D(h)): \\ h \in \mathcal{H} \cup \mathcal{H}_{\text{trivial}} \end{array} \right\}$

ROC

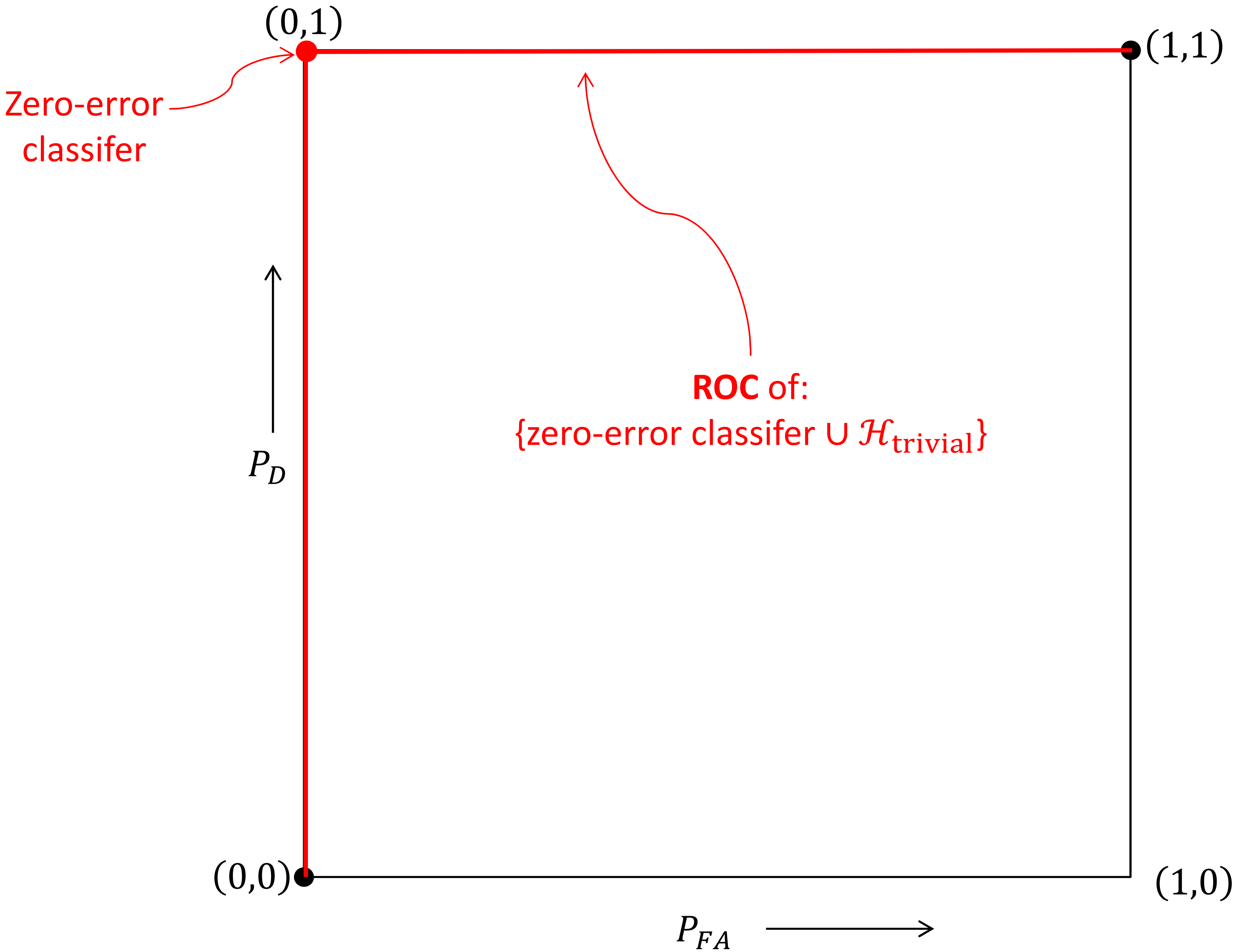
ROC = Upper envelope of convex hull of:
$\{(P_{FA}(h), P_D(h)) : h \in \mathcal{H} \cup \mathcal{H}_{\text{trivial}}\}$



Receiver Operating Characteristic (ROC)

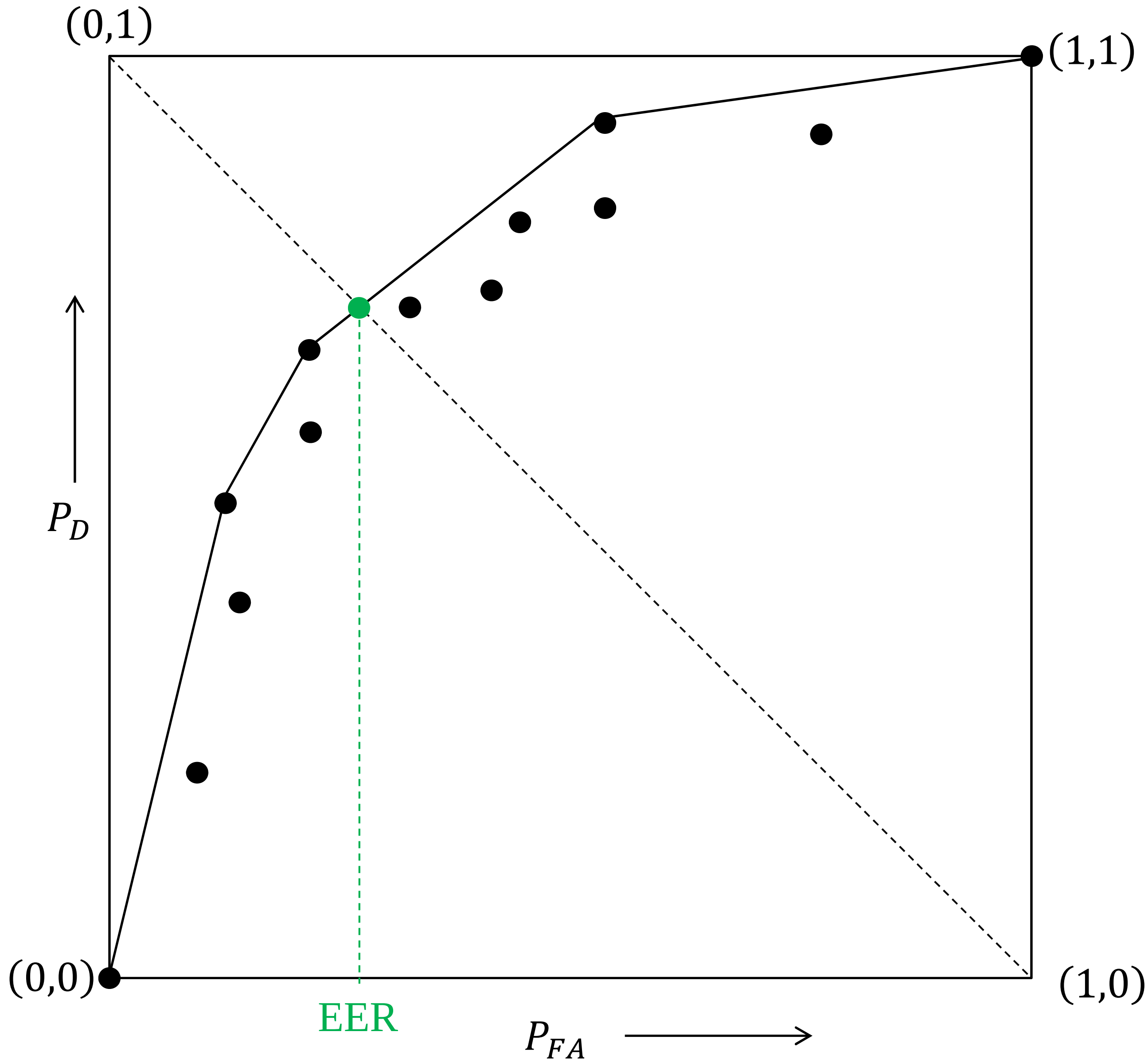
- A classifier that can perfectly separate positives from negatives (if one exists) will have an ROC curve which "hugs" the left-vertical and top horizontal axes (red curve in next figure). Such a classifier may not exist.
- The closer that an ROC curve of a family of classifiers is to "hugging" the ROC curve of the zero-error classifier, the better it is.
- The overall quality of an ROC curve is sometimes summarized as a single number:
  - Area Under the Curve (AUC): higher is better. Maximum is 1.
  - Equal Error Rate (EER) or Cross Over Rate: value of P_{FA} when $1 - P_D = P_{FA}$. Lower is better. Minimum is zero.

ROC

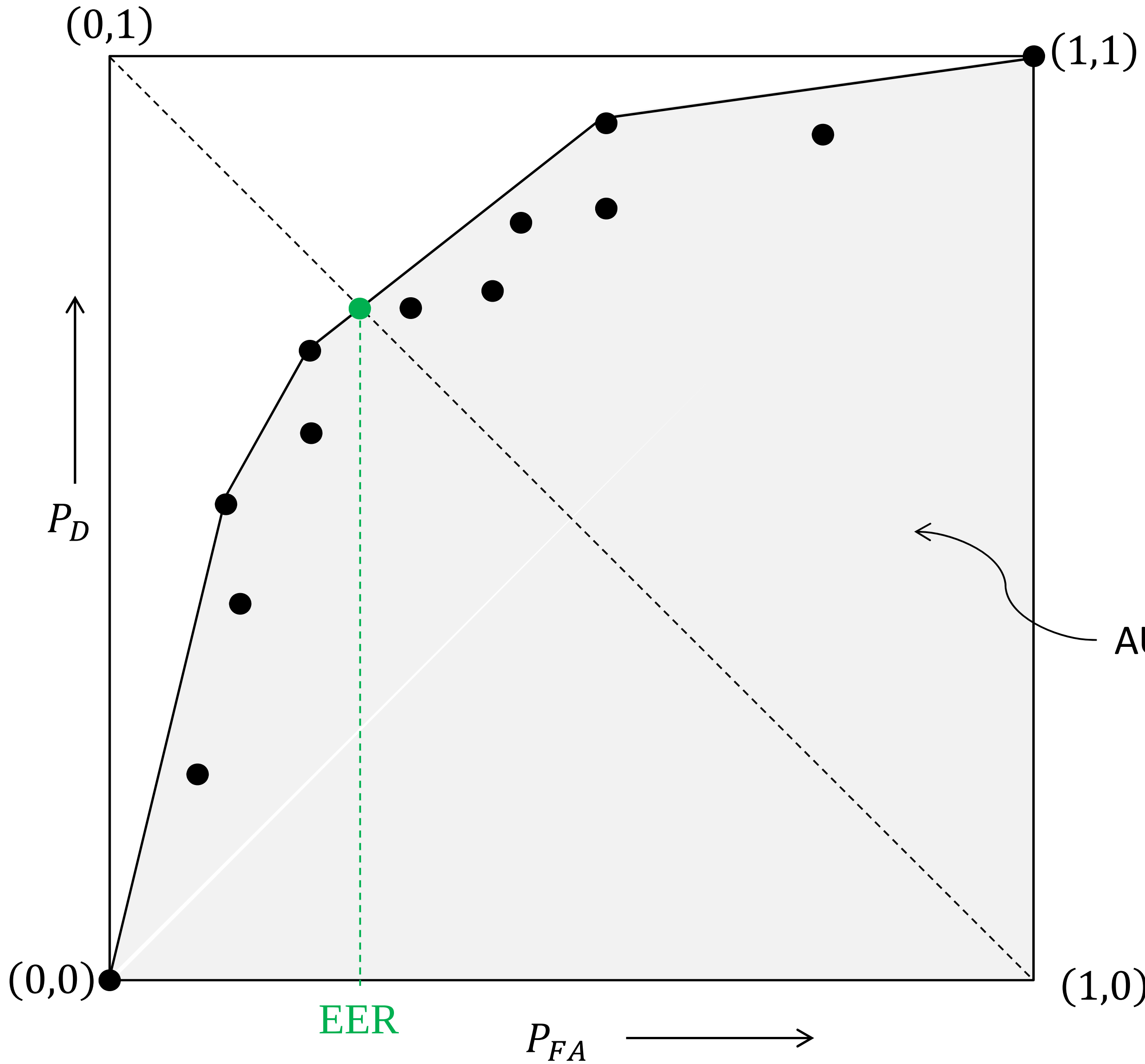


(0,1)
(1,1)
Zero-error classifer
ROC of:
{zero-error classifer ∪ $\mathcal{H}_{\text{trivial}}$}
P_D
(0,0)
(1,0)
P_{FA}

ROC



ROC



Receiver Operating Characteristic (ROC)

- Some properties of any ROC curve:
  - Always includes the points (0,0) and (1,1)
  - Always above the ROC curve of the trivial rules where $P_D = P_{FA}$
  - Shape of ROC curve is always concave: the chord joining any two points on the ROC curve is never above the ROC curve
  - ROC curve is always continuous, but need not be differentiable: e.g., can have consecutive straight line segments of different slopes
- In practice: (P_{FA}, P_D) estimated by (FPR, TPR)
  - i.e., empirical estimates of false-alarm and detection probabilities of different rules in the family
  - ideally, one should also show confidence bands in both directions around each point: (FPR$\pm k\hat{\sigma}_{\text{FPR}}$,TPR$\pm k\hat{\sigma}_{\text{TPR}}$) where k = 1 for 68% confidence, k = 2 for 95% confidence and k = 3 for 99% confidence

Likelihood Ratio Tests and ROC

- Likelihood Ratio (LR):

$$LR(\mathbf{x}) = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)}$$

- Randomized Likelihood Ratio Test (LRT) with threshold η and randomization probability γ:

$$h_{LRT}(\mathbf{x}) = \begin{cases} 1 \text{ with probability } 1 & \text{if } LR(\mathbf{x}) > \eta \\ 1 \text{ with probability } \gamma & \text{if } LR(\mathbf{x}) = \eta \\ 0 \text{ with probability } 1 & \text{if } LR(\mathbf{x}) < \eta \end{cases}$$

Likelihood Ratio Tests and ROC

- The family of randomized LRTs is optimal in the Neyman-Pearson sense:
  - Any $\mathrm{LRT}_{\eta,\gamma}$ with parameters η, γ is the most powerful test of its size
  - Meaning: if any decision rule h has a better (lower) false alarm probability than then $\mathrm{LRT}_{\eta,\gamma}$ it must have a worse (lower) detection probability:
  - Formally, if size = $P_{FA}(h) \leq P_{FA}(\mathrm{LRT}_{\eta,\gamma})$ then power = $P_D(h) \leq P_D(\mathrm{LRT}_{\eta,\gamma})$

- If $\gamma = 1$ and $\eta \to \infty$, then $\mathrm{LRT}_{\eta,\gamma} \to h_1$ the trivial "always decide class one" rule

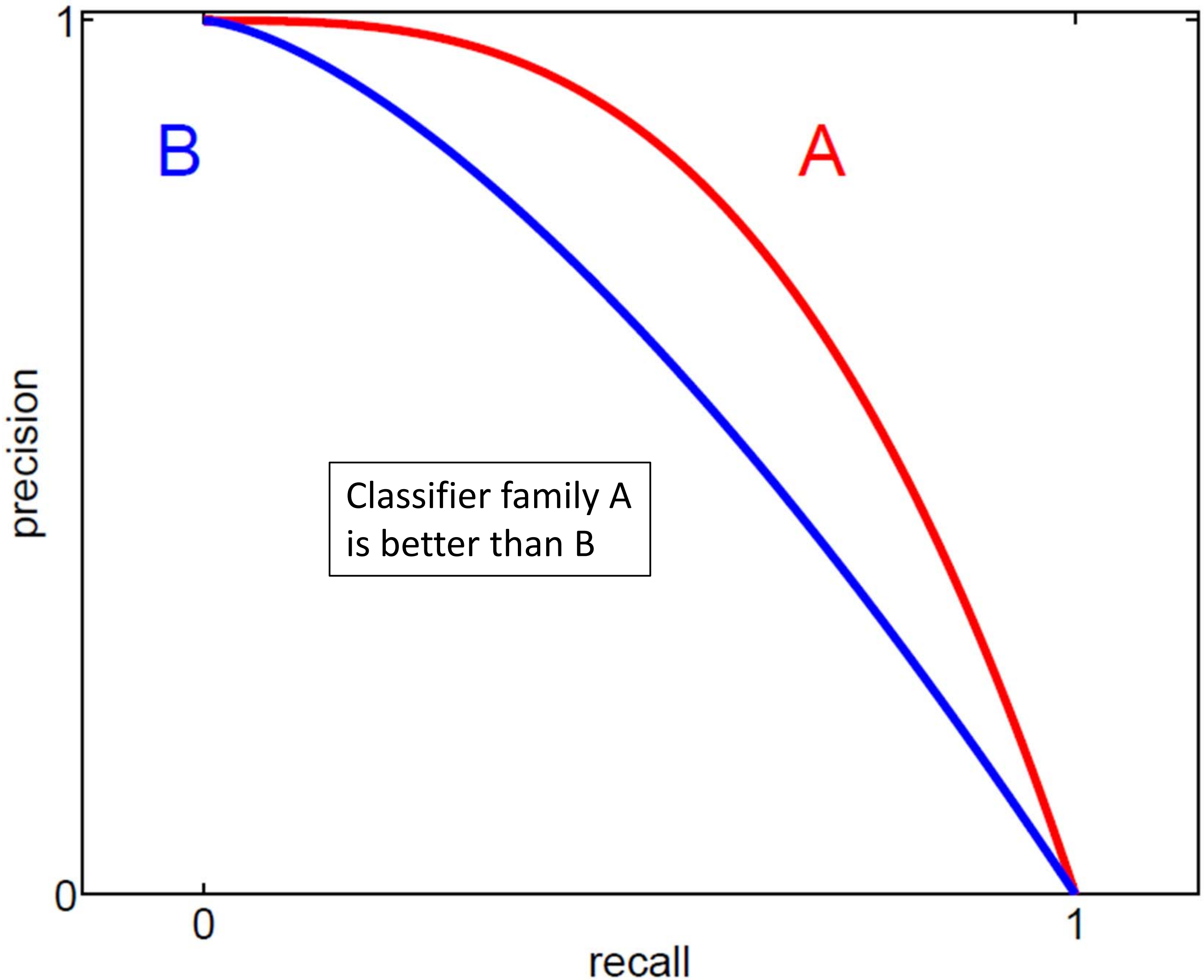
- If $\gamma = 0$ and $\eta \to 0$, then $\mathrm{LRT}_{\eta,\gamma} \to h_0$ the trivial "always decide class zero" rule

- If the ROC curve of the family of LRTs is differentiable at a point, then its slope at that point = the LRT-threshold = η

Precision-Recall Curve

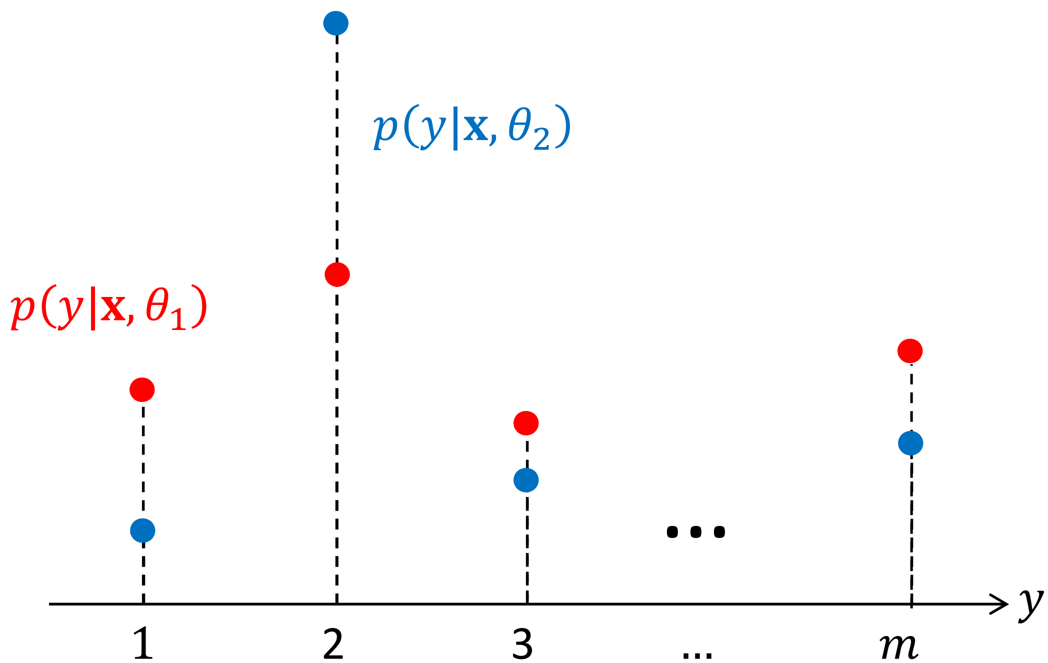
- Similar in concept to ROC, except it is a plot of precision versus recall (see figure).
- Here, "hugging" the top right is the best we can do.
- This curve can be summarized by a single number:
  - The mean precision (averaged over recall values) which approximates the area under the curve.
  - Alternatively, one can quote the precision for a recall, e.g., when the first $K = 10$ positive entries have been correctly recalled. This is called average precision at K score. Used in evaluating information retrieval systems.

Precision-Recall Curve



log-loss, perplexity

- Consider the following 2 choices for the posterior pmf:



- Both will produce the same MAP decision (2 in this example), but clearly model $p(y|\mathbf{x},\theta_2)$ is more "decisive" than model $p(y|\mathbf{x},\theta_1)$ for this $\mathbf{x}$

- ⇒Need a more nuanced test sample performance measure than CCR to tease-apart the "decisiveness" of different models/estimates of the posterior pmf of the label given the feature

log-loss, perplexity

$$\text{logloss} = -\frac{1}{n_{\text{test}}}\sum_{j=1}^{n_{\text{test}}} \log_b p(y_j|\mathbf{x}_j,\theta), \quad \text{typically, } b=2 \text{ for “bits”}.$$

$$\text{perplexity} = b^{\text{logloss}}$$

- If test data $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n_{\text{test}}}, Y_{n_{\text{test}}})$are ~ IID samples from a joint distribution $q(\mathbf{x}, y)$ then, as $n_{\text{test}} \to \infty$, by the law of large numbers, the logloss will converge to:

$$\text{logloss} \to E_{(\mathbf{X},Y)\sim q(\mathbf{x},y)}[-\log_b p(Y|\mathbf{X},\theta)]$$

- This limiting logloss can be shown to be the sum of the following 2 nonnegative terms:

$$E_{(\mathbf{X},Y)\sim q(\mathbf{x},y)}[-\log_b p(Y|\mathbf{X},\theta)] = \underbrace{E_{(\mathbf{X},Y)\sim q(\mathbf{x},y)}[-\log_b q(Y|\mathbf{X})]}_{\text{Conditional entropy of label given feature}} + \underbrace{E_{\mathbf{X}\sim q(\mathbf{x})}[D(q(y|\mathbf{X})||p(y|\mathbf{X},\theta))]}_{\text{Conditional divergence between true posterior and model}}$$

- See Problem 2.4 in Assignment 2 for definition of divergence $D(q||p)$ between 2 pmfs