Bayes Decision Theory

Chenxi Liu
2018/09/20

Guessing a lady’s age

- You asked a girl “What’s your age?”
- She said “What’s your guess?”
- Somehow you have narrowed down to either 20 or 30. Which one should you answer?

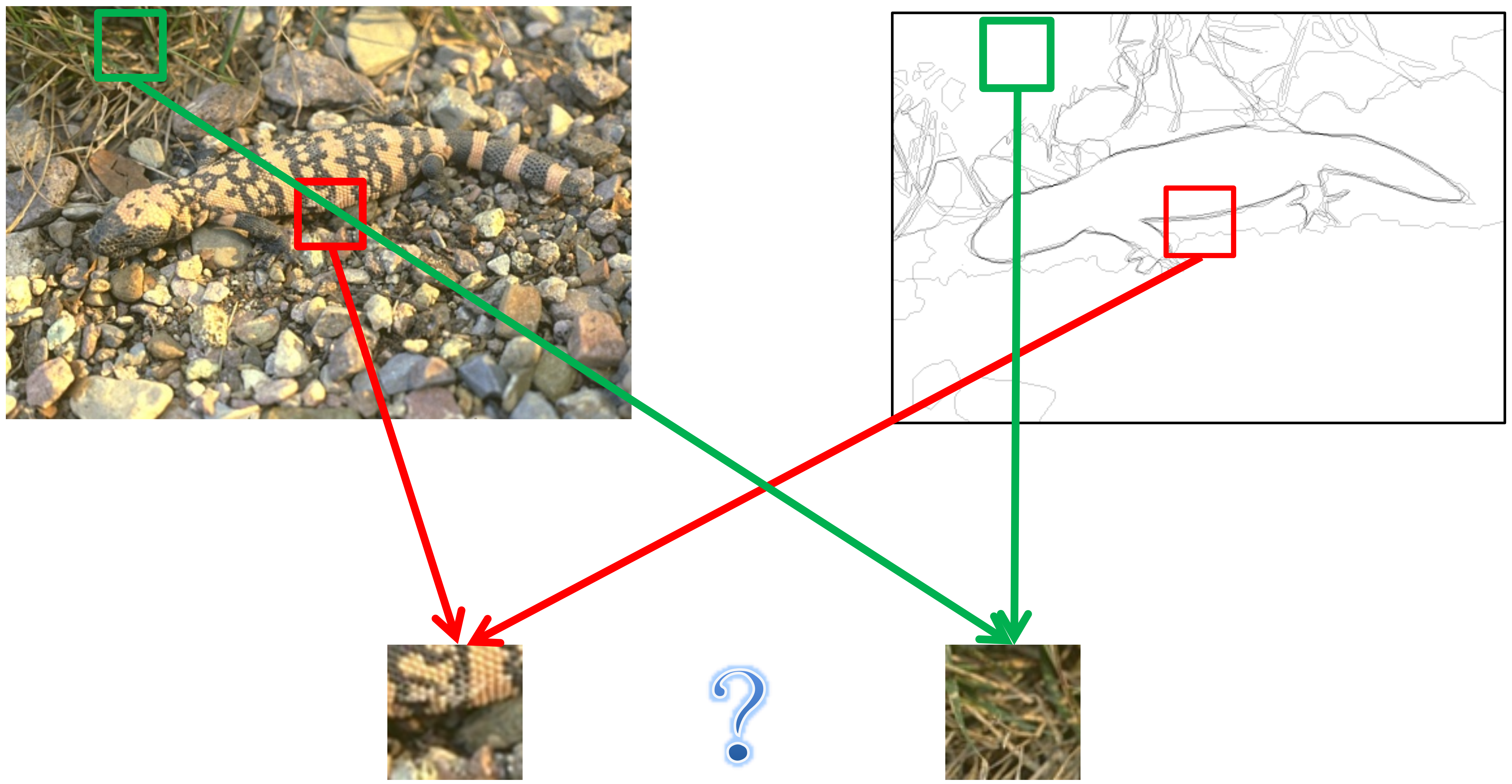
What factors help you decide?

- Cue 1: Appearance (for simplicity, just consider eye size as single feature)
- Cue 2: Prior knowledge about age distribution
- Cue 3: Reward/penalty if you got the answer correct/wrong

Edge detection

- For every pixel in the image, you would like to classify if it is edge or not

Edge detection



changes correspond to object boundaries

changes caused by texture patterns textures

What factors help you decide?

- Cue 1: Appearance (filter response discussed in last lecture)
- Cue 2: Prior knowledge about edge percentage
- Cue 3: Reward/penalty if you got the answer correct/wrong

Same framework for both

- These two scenarios are actually very similar in nature: we want to make the “optimal” decision!
- **Bayes decision theory** is a framework for making optimal decisions in the presence of uncertainty

Notations

- Input: $x \in \mathcal{X}$ (e.g. features/filter responses of the image)
- Output: $y \in \mathcal{Y}$ (e.g. +1 for 20 years old/edge is present; -1 for 30 years old/edge is not present)
- A probability distribution $P(x, y)$ generates the input and output
- A decision rule $\hat{y} = \alpha(x)$
- Loss function $L(\alpha(x), y)$ captures the cost of making decision $\alpha(x)$ if the real answer is y

Notations

- The risk is specified by $R(\alpha) = \sum_{x,y} P(x, y)L(\alpha(x), y)$
- The Bayes rule is $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$
- The Bayes risk is $\min_{\alpha} R(\alpha) = R(\hat{\alpha})$

Deriving Bayes decision rule

- Usually we don't have the explicit distribution $P(x)$; instead, we have a limited number of samples sampled from this distribution

- Therefore, in practice we minimize the following empirical risk:

$$\hat{R}(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\sum_{y} P(y\,|\,x_i)L(\alpha(x_i), y) \quad x_i \sim P(x)$$

Deriving Bayes decision rule

- This basically means for every x_i that we see, we wish to minimize the risk it invokes $\sum_y P(y|x_i)L(\alpha(x_i), y)$
- Therefore, the "optimal" decision rule is:

$$\hat{\alpha}(x_i) = \arg\min_{\alpha} \sum_y P(y|x_i)L(\alpha(x_i), y) = \arg\min_{\alpha} \sum_y P(x_i|y)P(y)L(\alpha(x_i), y)$$

Binary decision problems

- Four possibilities:

$$L(\alpha(x) = 1, y = 1) = T_p \qquad L(\alpha(x) = -1, y = 1) = F_n$$

$$L(\alpha(x) = 1, y = -1) = F_p \quad L(\alpha(x) = -1, y = -1) = T_n$$

- The expected "risk" of predicting 1: $T_p P(y = 1 | x) + F_p P(y = -1 | x)$

- The expected "risk" of predicting -1: $F_n P(y = 1 | x) + T_n P(y = -1 | x)$

Binary decision problems

- We should predict 1 instead of -1 when its expected "risk" is smaller:

$$T_p P(y = 1 | x) + F_p P(y = -1 | x) < F_n P(y = 1 | x) + T_n P(y = -1 | x)$$

$$(F_n - T_p) P(y = 1 | x) > (F_p - T_n) P(y = -1 | x)$$

$$\frac{P(y = 1 | x)}{P(y = -1 | x)} > \frac{F_p - T_n}{F_n - T_p}$$

$$\frac{P(x | y = 1) P(y = 1)}{P(x | y = -1) P(y = -1)} > \frac{T_n - F_p}{T_p - F_n}$$

$$\frac{P(x | y = 1)}{P(x | y = -1)} > \frac{T_n - F_p}{T_p - F_n} \frac{P(y = -1)}{P(y = 1)}$$

Binary decision problems

- Log-likelihood ratio test:

$$\log \frac{P(x \mid y = 1)}{P(x \mid y = -1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y = -1)}{P(y = 1)}$$

- The intuition is that the evidence in the log-likelihood must be bigger than our prior biases while taking into account the penalties paid for different types of mistakes

Guessing a lady’s age

- Cue 1: Appearance
    - Given she is 20 years old (+1), the probability of the observed eye size is 30%; Given she is 30 years old (-1), this probability is 20%.
- Cue 2: Prior knowledge about age distribution
    - Suppose there was a baby boom 30 years ago; so in the current female population, 30% are age 30 and only 20% are age 20.

Guessing a lady's age

- Cue 3: Reward/penalty if you got the answer correct/wrong
  - If you guessed right, perfect, no hard feelings
  - If you guessed 20 and the truth is 30, you pay a small cost
  - If you guessed 30 and the truth is 20, you pay a BIG cost

Guessing a lady's age

- Recall: you should predict 1 (20 years old) instead of -1 (30 years old) when the following holds:

$$\frac{P(x \mid y = 1)}{P(x \mid y = -1)} > \frac{T_n - F_p}{T_p - F_n} \frac{P(y = -1)}{P(y = 1)}$$

- Indeed,

$$\frac{0.3}{0.2} > \frac{0 - 1}{0 - 100} \frac{0.3}{0.2}$$

Special case 1: MAP

- If the loss function penalizes all errors by the same amount,

$$L(\alpha(x), y) = K_1 \quad \alpha(x) \neq y$$

$$L(\alpha(x), y) = K_2 \quad \alpha(x) = y$$

- then the Bayes rule corresponds to the maximum a posteriori estimator

$$\alpha(x) = \arg\max_y P(y \mid x)$$

Special case 1: MAP

- In binary decision problems, this means we should predict 1 instead of -1 when

$$\frac{P(y=1|x)}{P(y=-1|x)} > \frac{F_p - T_n}{F_n - T_p} = 1$$

- In n-class setting, if $K_1 = 1, K_2 = 0$, then the “risk” of choosing class j is

$$\sum_y P(y|x)L(\alpha(x), y) = \sum_{y \neq j} P(y|x) = 1 - P(y=j|x)$$

$$\alpha(x) = \arg\min_y (1 - P(y|x)) = \arg\max_y P(y|x)$$

Special case 2: MLE

- If, in addition, the prior is a uniform distribution,

$$P(y) = C \quad \forall y$$

- then Bayes rule reduces to the maximum likelihood estimate

$$\alpha(x) = \arg\max_{y} P(x \mid y)$$

Edge detection

- We have derived that we should predict a pixel is edge (1) instead of non-edge (-1) when

$$\log \frac{P(x \mid y = 1)}{P(x \mid y = -1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y = -1)}{P(y = 1)} = T$$

Edge detection

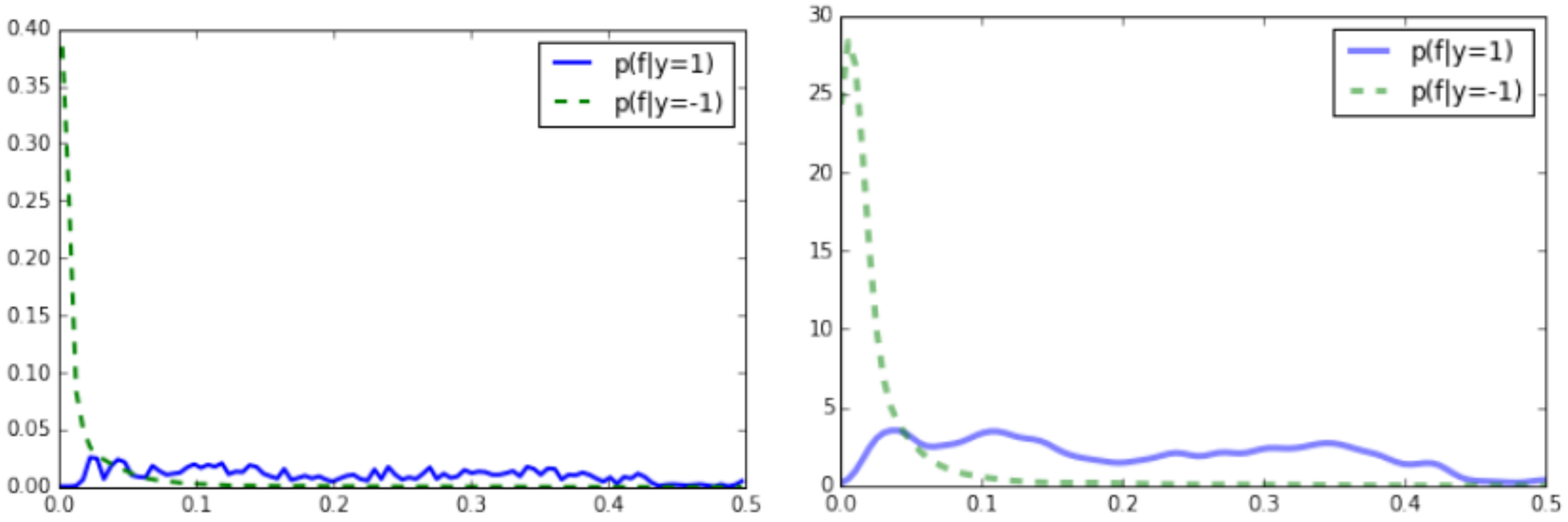


Figure 21 : The probability of filter responses conditioned on whether the filter is *on* or *off* an edge – $P(f|y=1), P(f|y=-1)$, where $f(x) = |\vec{\nabla} I(x)|$. Left: The probability distributions learned from a data set of images. Right: The smoothed distributions after fitting the data to a parametric model.

Edge detection

- We have derived that we should predict a pixel is edge (1) instead of non-edge (-1) when

$$\log \frac{P(x \mid y=1)}{P(x \mid y=-1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y=-1)}{P(y=1)} = T$$

- But what if we don't want to pick exact values for penalties T_n, F_p, T_p, F_n ?

Edge detection

$$\log \frac{P(x \mid y = 1)}{P(x \mid y = -1)} > T$$

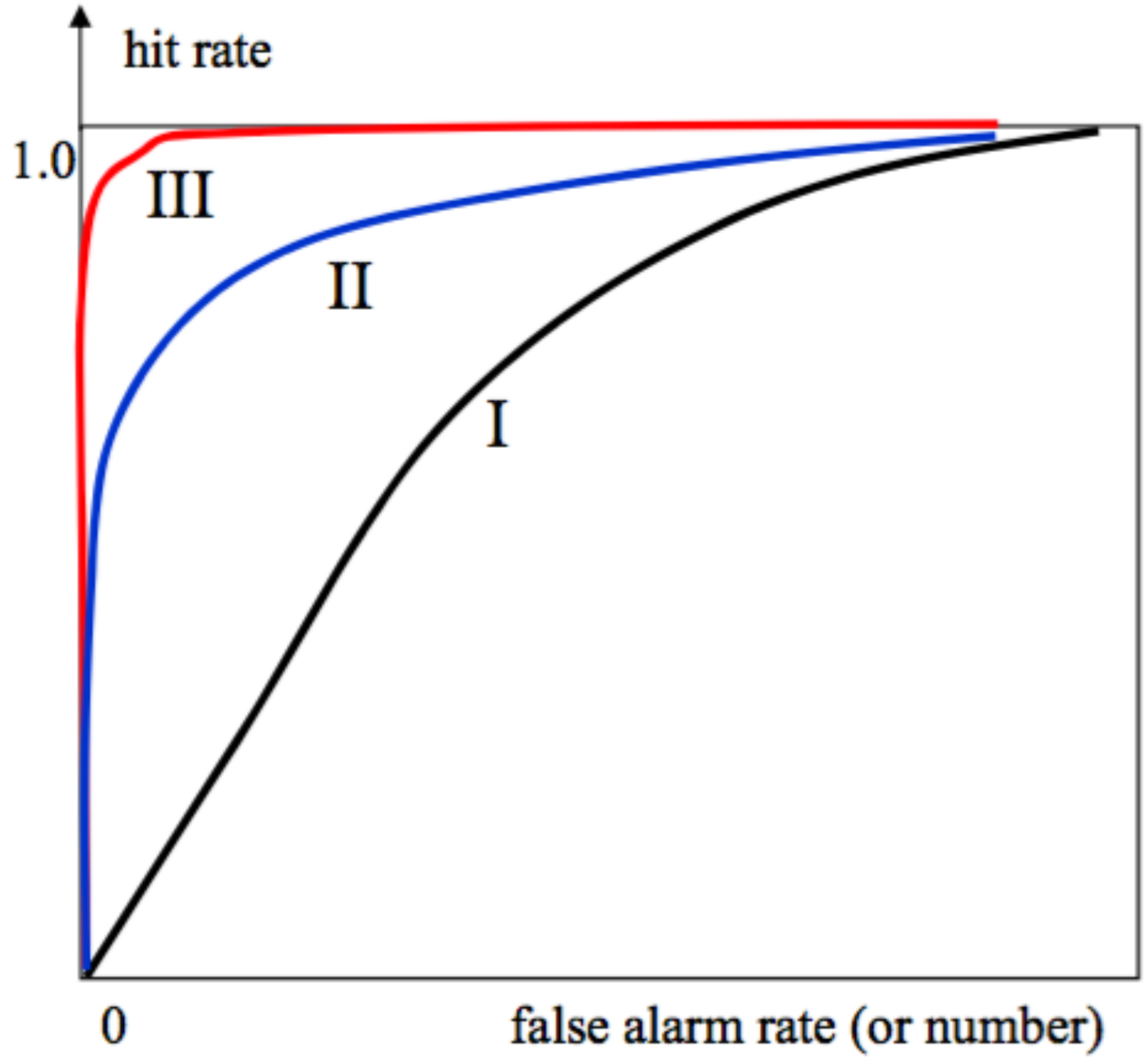
- When the threshold is small:
  - Very easy to predict pixel as edge
  - High true positive rate (close to 1); High false positive rate (close to 1)
- When the threshold is large:
  - Very hard to predict pixel as edge
  - Low true positive rate (close to 0); Low false positive rate (close to 0)

ROC curve

- The receiver operating characteristic (ROC) curve tries to capture this trade-off between true positive rate and false positive rate

- Which point corresponds to very small/ large threshold?

- Which curve is the best?



Take-home messages

- You probably already knew it is wise to guess a younger age…
- But now you can explain your action under Bayes decision theory!
- And pretty much the same thing goes on for edge detection and a lot other computer vision and machine learning tasks
- We have mostly focused on binary classification, but straightforward extensions exist for multi-way classification