Classification metrics

Plan for the video

- Accuracy
- Logarithmic loss
- Area under ROC curve
- (Quadratic weighted) Kappa

Notation

- N – is number of objects
- L – is number of classes

- y – ground truth
- $\hat{y}$ – predictions
- $[a = b]$ – indicator function

- Soft labels (soft predictions) are classifier's scores

- Hard labels (hard predictions):
  - $\arg\max_i f_i(x)$
  - $[f(x) > b]$, b – threshold

Accuracy score

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i]$$

- How frequently our class prediction is correct.

Accuracy score

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\alpha = y_i]$$

- How frequently our class prediction is correct.
- Best constant:
  - **predict the most frequent class.**

Accuracy score

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\alpha = y_i]$$

- How frequently our class prediction is correct.
- Best constant:
  - **predict the most frequent class.**

- Dataset:
  - 10 cats
  - 90 dogs

Predict always dog:

Accuracy = **0.9**!

Logarithmic loss (logloss)

- Binary:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

$$y_i \in \mathbb{R}, \quad \hat{y}_i \in \mathbb{R}$$

Logarithmic loss (logloss)

- Binary:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

$$y_i \in \mathbb{R}, \quad \hat{y}_i \in \mathbb{R}$$

- Multiclass:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{L} y_{il} \log(\hat{y}_{il})$$

$$y_i \in \mathbb{R}^L, \quad \hat{y}_i \in \mathbb{R}^L$$

Logarithmic loss (logloss)

- Binary:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$$y_i \in \mathbb{R}, \quad \hat{y}_i \in \mathbb{R}$$

- Multiclass:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{L} y_{il}\log(\hat{y}_{il})$$

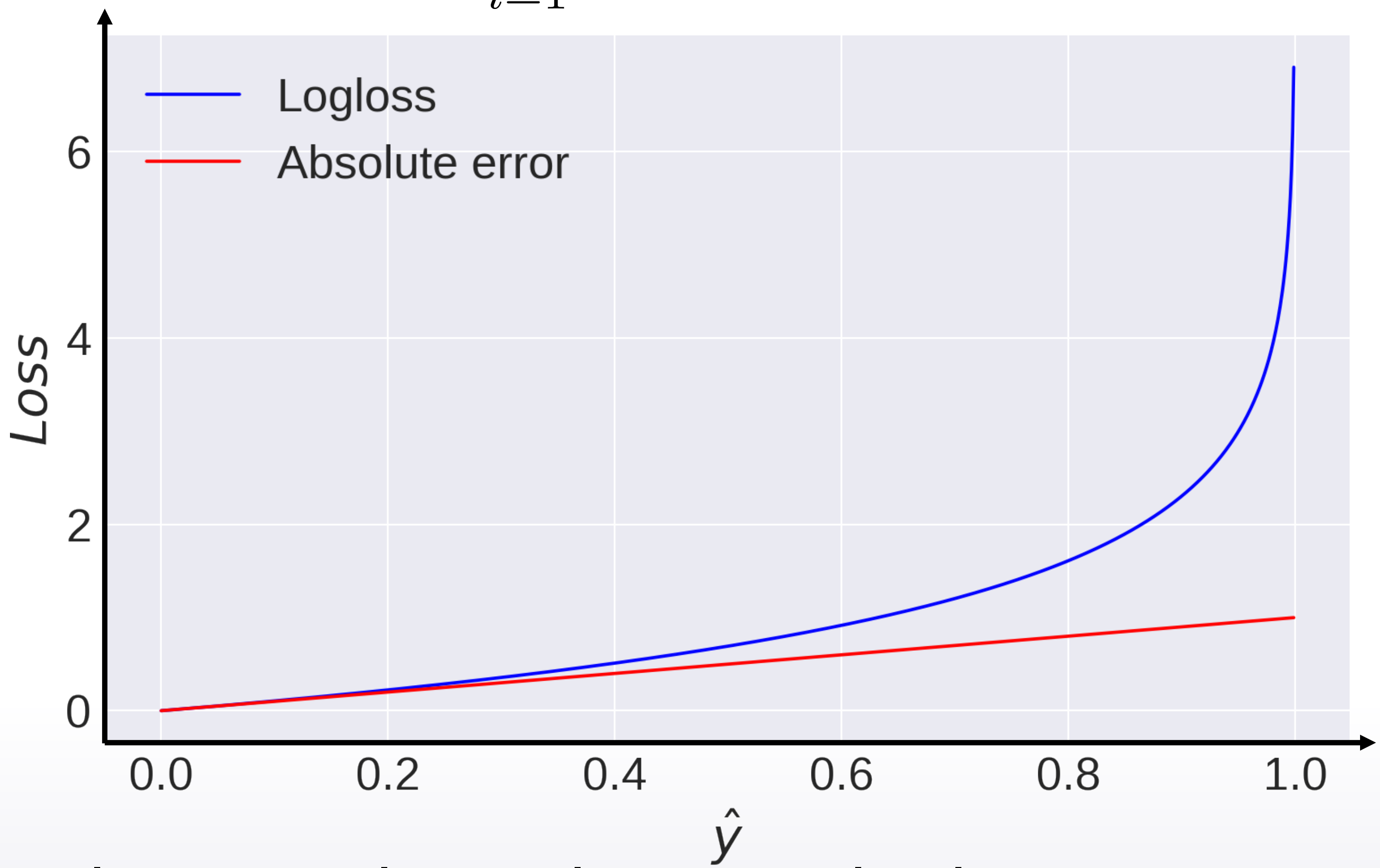
$$y_i \in \mathbb{R}^L, \quad \hat{y}_i \in \mathbb{R}^L$$

- In practice:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{L} y_{il}\log(\min(\max(\hat{y}_{il}, 10^{-15}), 1-10^{-15}))$$

Logarithmic loss (logloss)

$$\mathrm{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$



- Logloss strongly penalizes completely wrong answers

Logarithmic loss (logloss)

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\alpha) + (1 - y_i)\log(1 - \alpha)$$

- Best constant:
  - **set α_i to frequency of *i*-th class.**

- `Dataset:`
  - `10 cats`
  - `90 dogs`

$$\alpha = [0.1, 0.9]$$

Area Under Curve (AUC ROC)

$$\mathrm{Accuracy}([\hat{y} > 0.5]) = \frac{6}{7}$$

Area Under Curve (AUC ROC)

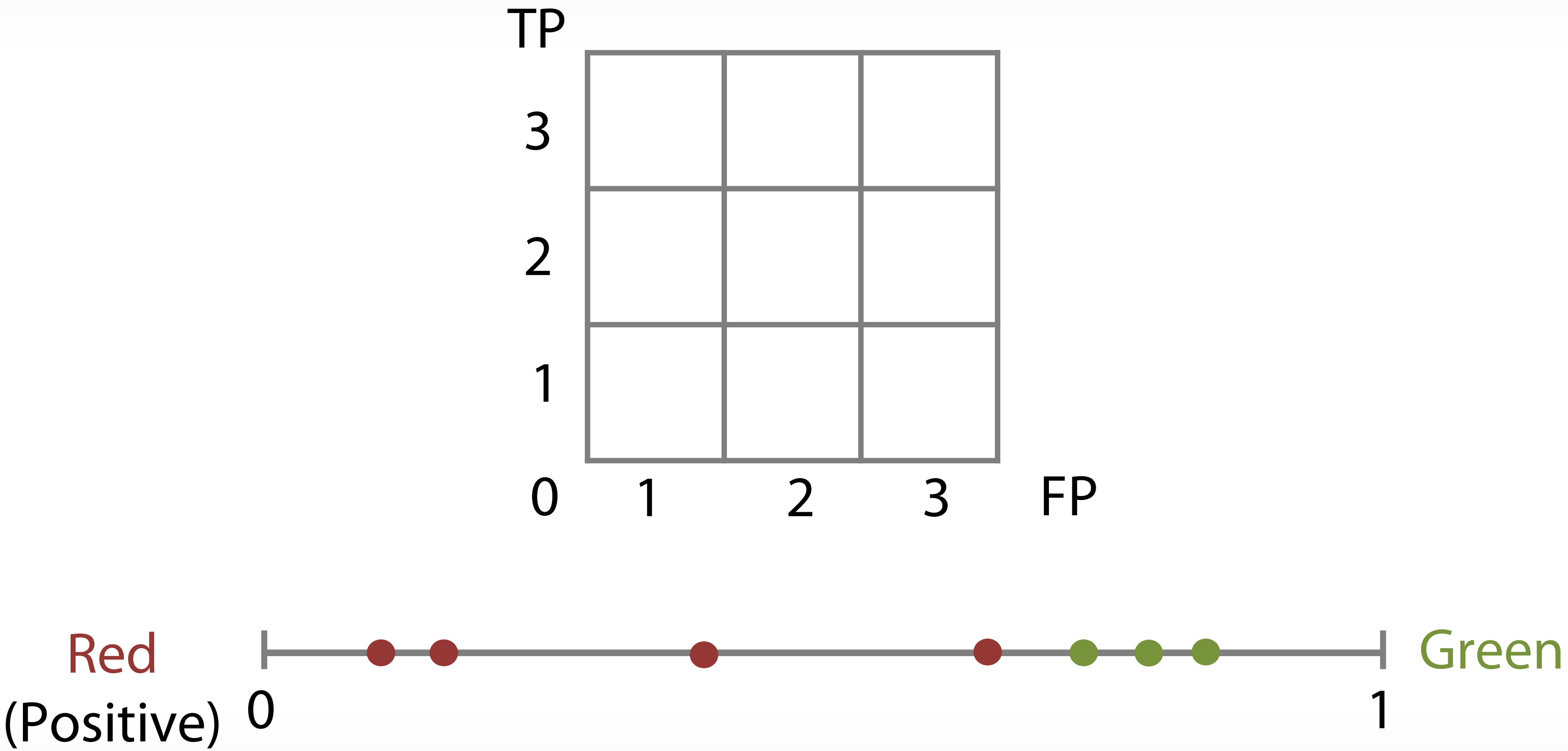
$$\text{Accuracy}([\hat{y} > 0.7]) = 1$$

Area Under Curve (AUC ROC)

$$\text{Accuracy}([\hat{y} > {\color{blue}0.7}]) = {\color{red}1}$$

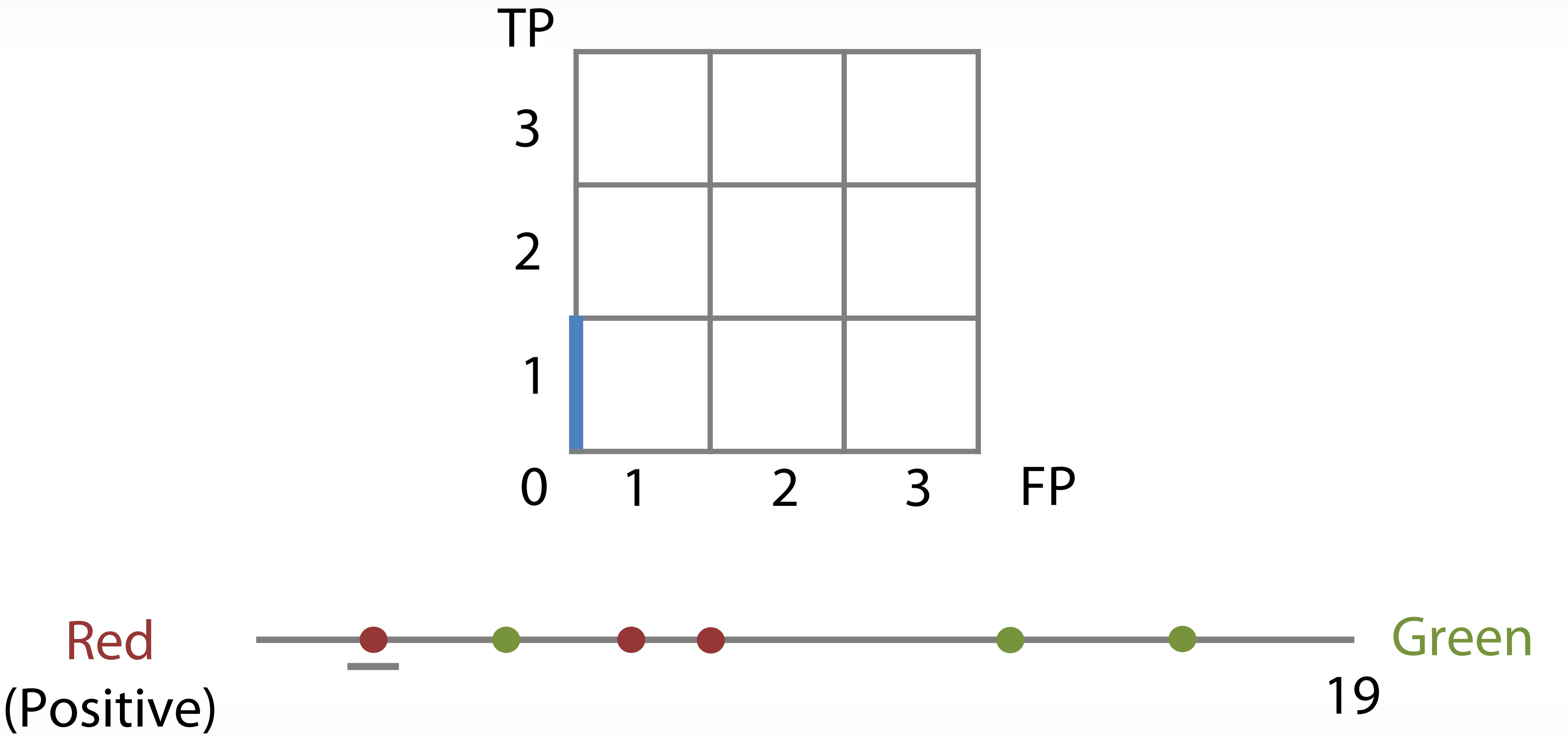
- Only for binary tasks
- Depends only on ordering of the predictions, not on absolute values
- **Several explanations**
  1) Area under curve
  2) Pairs ordering

Area Under Curve (AUC ROC)



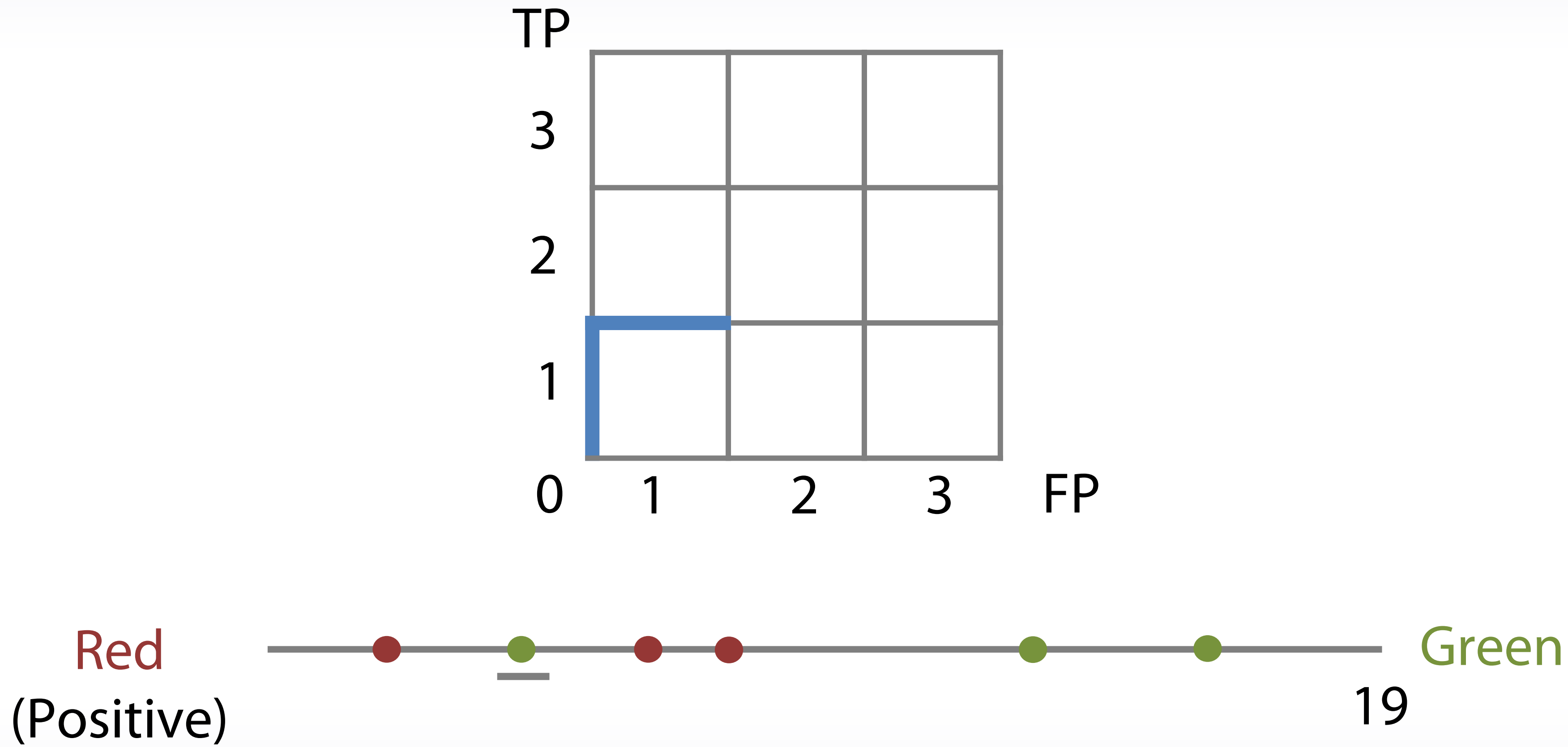
TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)



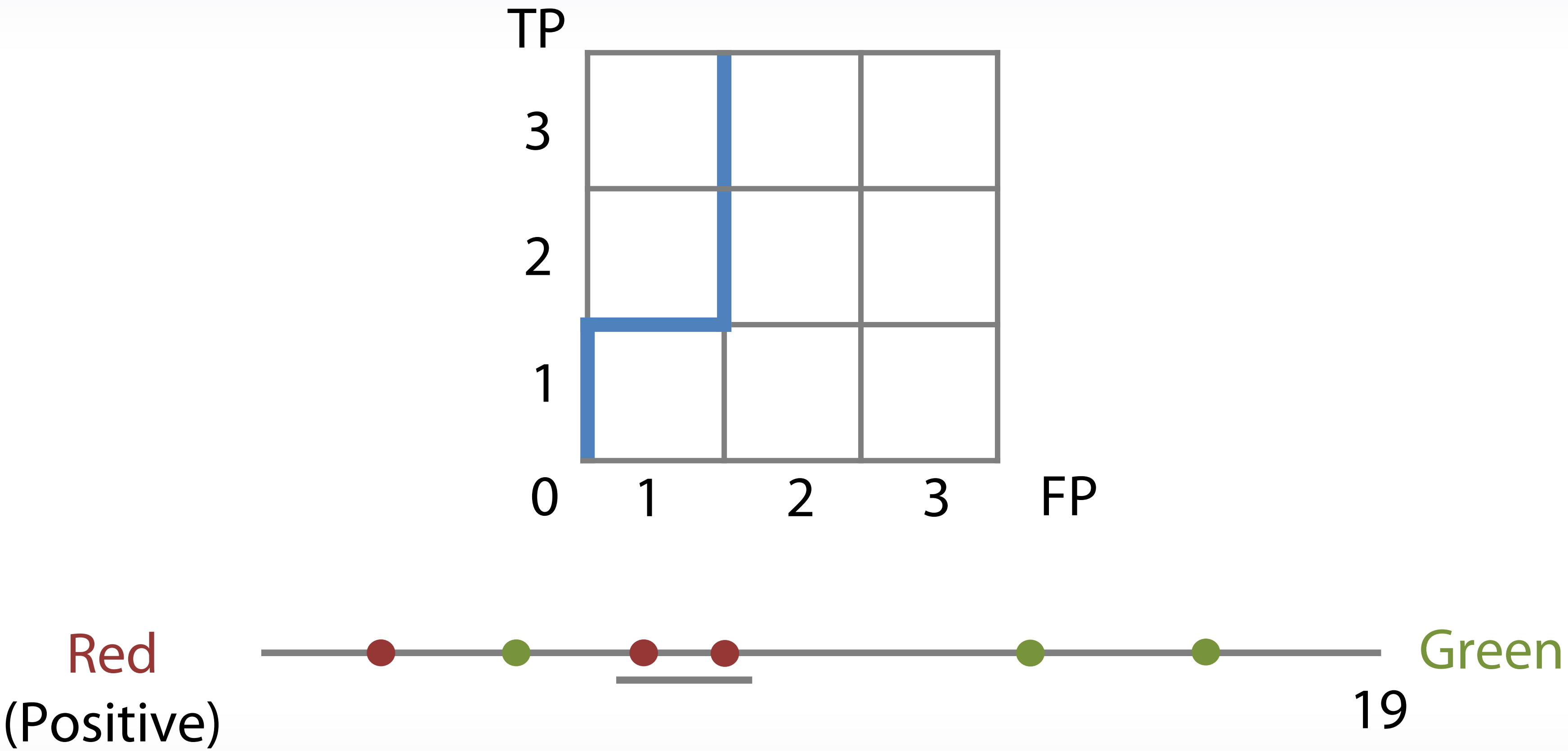
TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)



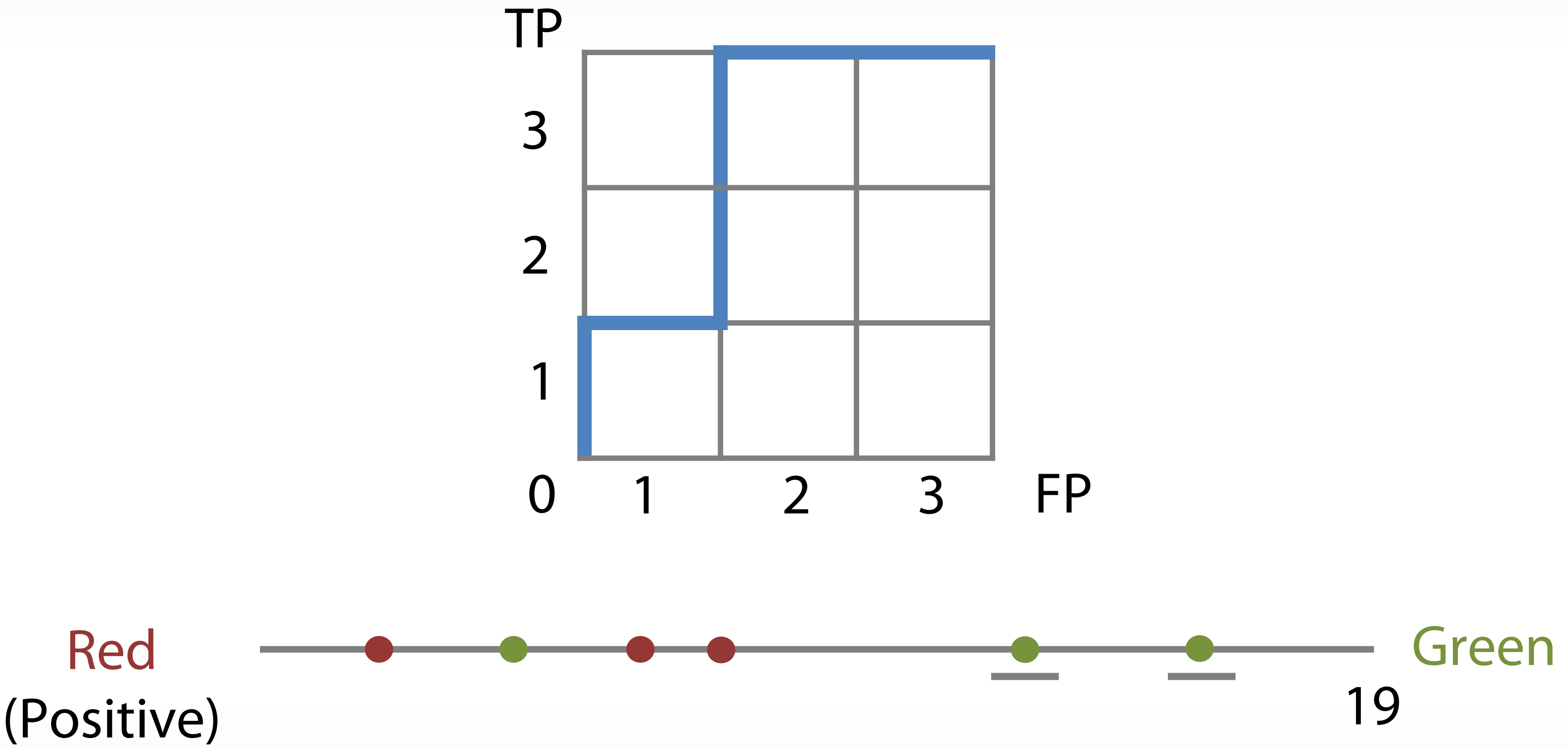
TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)



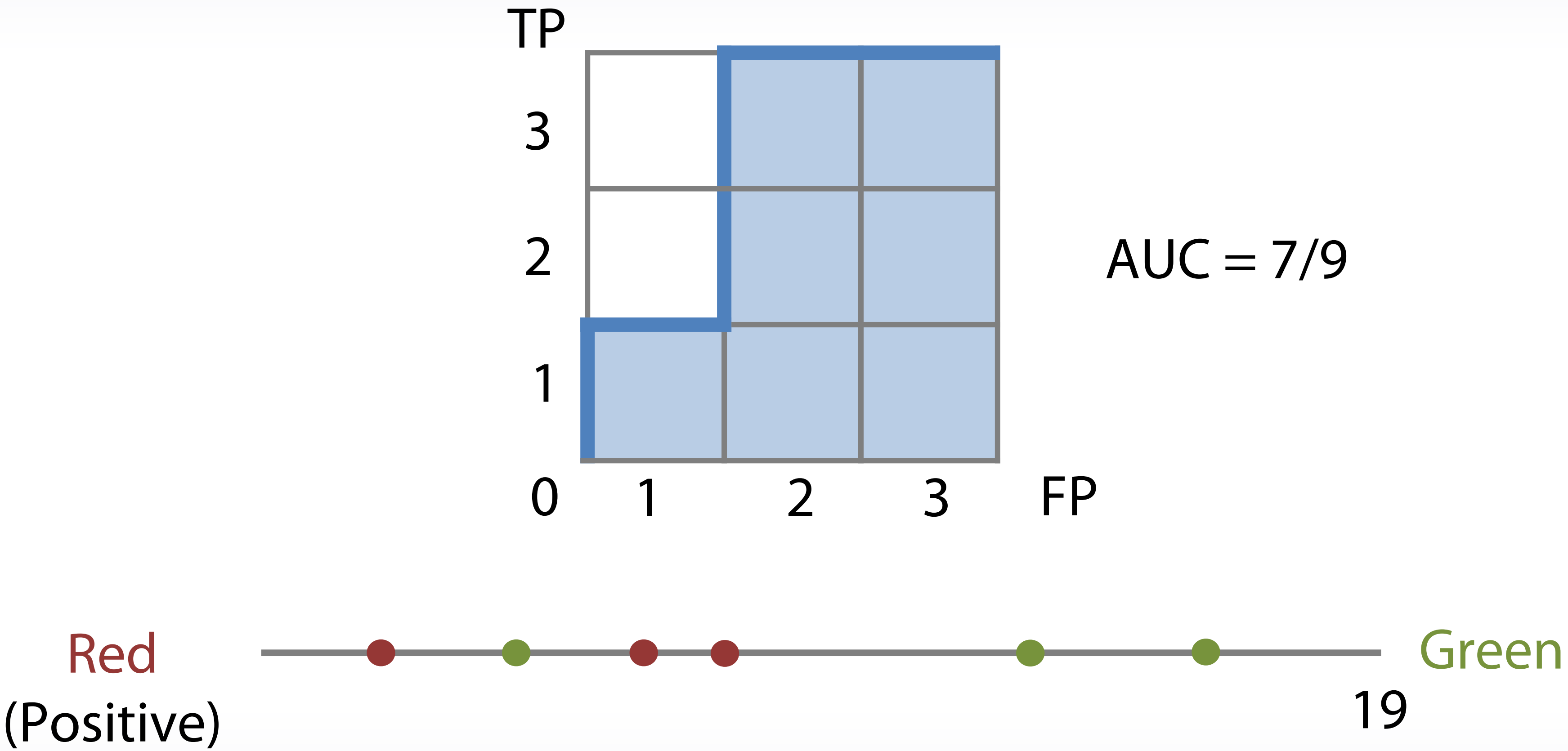
TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)



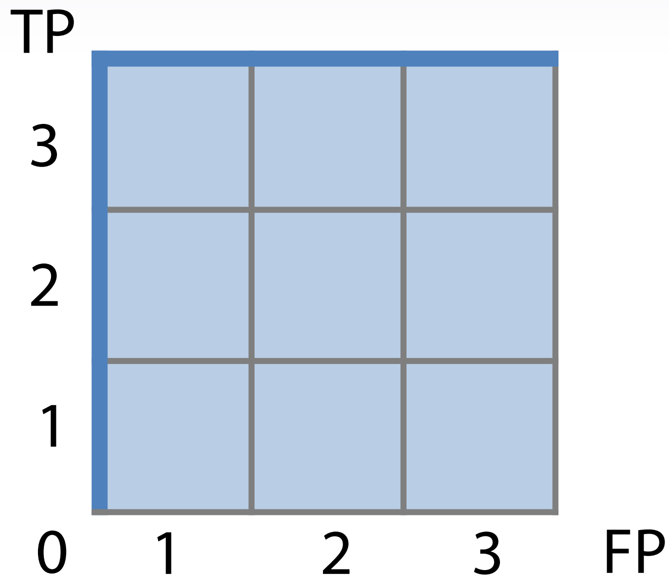
TP – true positives, FP – false positives

Area Under Curve (AUC ROC)

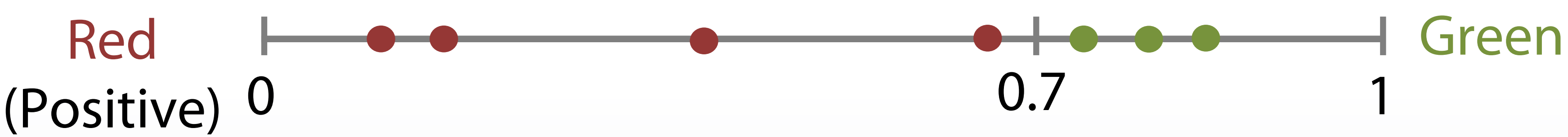


TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)

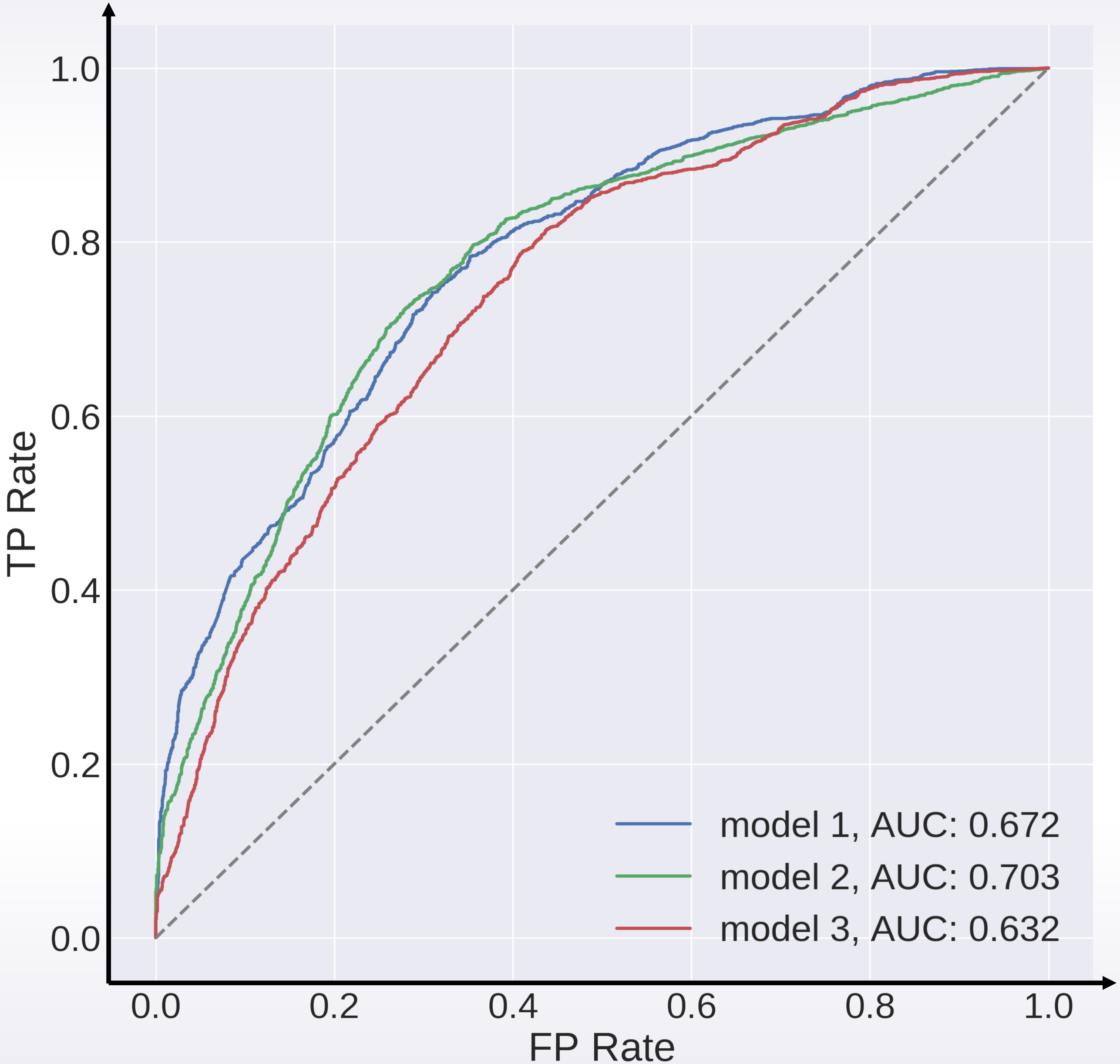


AUC = 1



TP – true positives, **FP** – false positives

Area Under Curve (AUC ROC)



Area Under Curve (AUC ROC)

Area Under Curve (AUC ROC)

$$\mathrm{AUC} = \frac{\#\ \text{correctly ordered pairs}}{\text{total number of pairs}} = $$
$$= 1 - \frac{\#\ \text{incorrectly ordered pairs}}{\text{total number of pairs}}$$

pair = (red object, green object)

Area Under Curve (AUC ROC)

- **Best constant:**
  - All constants give same score

- **Random predictions lead to AUC = 0.5**

Cohen's Kappa motivation

Dataset:

- 10 cats
- 90 dogs

Baseline accuracy = 0.9

$$\text{my\_score} = 1 - \frac{1 - \text{accuracy}}{1 - \text{baseline}}$$

- accuracy = 1 → my_score = 1
- accuracy = 0.9 → my_score = 0

Cohen's Kappa motivation

Dataset:

- 10 cats
- 90 dogs

Predict *20 cats* and *80 dogs* at random: *accuracy* ~ 0.74

0.2*0.1 + 0.8*0.9 = 0.74

$$\text{Cohen's Kappa} = 1 - \frac{1 - \text{accuracy}}{1 - p_e}$$

p_e – what accuracy would be on average, if we randomly permute our predictions

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

Cohen's Kappa motivation

Dataset:

- 10 cats
- 90 dogs

Predict *20 cats* and *80 dogs* at random: *accuracy* ~ 0.74

error ~ 0.26

$$\text{Cohen's Kappa} = 1 - \frac{\text{error}}{\text{baseline error}}$$

Weighted error

Dataset:
- 10 cats
- 90 dogs
- 20 tigers

Error weight matrix

pred\ true	cat	dog	tiger
cat	0	1	10
dog	1	0	10
tiger	1	1	0

Weighted error and weighted Kappa

Confusion matrix *C*

pred\ true	cat	dog	tiger
cat	4	2	3
dog	2	88	5
tiger	4	10	12

Weight matrix *W*

pred\ true	cat	dog	tiger
cat	0	1	10
dog	1	0	10
tiger	1	1	0

$$\text{weighted error} = \frac{1}{const} \sum_{i,j} C_{ij} W_{ij}$$

Weighted error and weighted Kappa

Confusion matrix *C*

pred\ true	cat	dog	tiger
cat	4	2	3
dog	2	88	4
tiger	4	10	12

Weight matrix *W*

pred\ true	cat	dog	tiger
cat	0	1	10
dog	1	0	10
tiger	1	1	0

$$\text{weighted error} = \frac{1}{const} \sum_{i,j} C_{ij} W_{ij}$$

$$\boxed{\text{weighted kappa} = 1 - \frac{\text{weighted error}}{\text{weighted baseline error}}}$$

Quadratic and Linear Weighted Kappa

Linear weights

pred\ true	**1**	**2**	**3**
1	0	1	2
2	1	0	1
3	2	1	0

$$w_{ij} = |i - j|$$

Quadratic weights

pred\ true	**1**	**2**	**3**
1	0	1	4
2	1	0	1
3	4	1	0

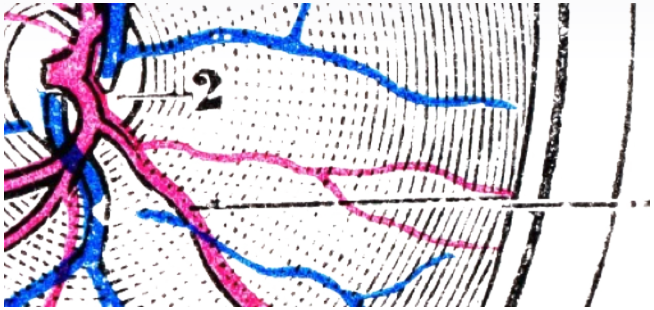
$$w_{ij} = (i - j)^2$$

$$\text{weighted kappa} = 1 - \frac{\text{weighted error}}{\text{weighted baseline error}}$$

Quadratic Weighted Kappa

CrowdFlower

CrowdFlower Search Results Relevance



Diabetic Retinopathy Detection

Prudential

Prudential Life Insurance Assessment

ASAP

Automated Student Assessment Prize

Phase One: Automated Essay Scoring

The Hewlett Foundation: Automated Essay Scoring

Conclusion

- Accuracy
- Logloss
- AUC (ROC)
- (Quadratic weighted) Kappa