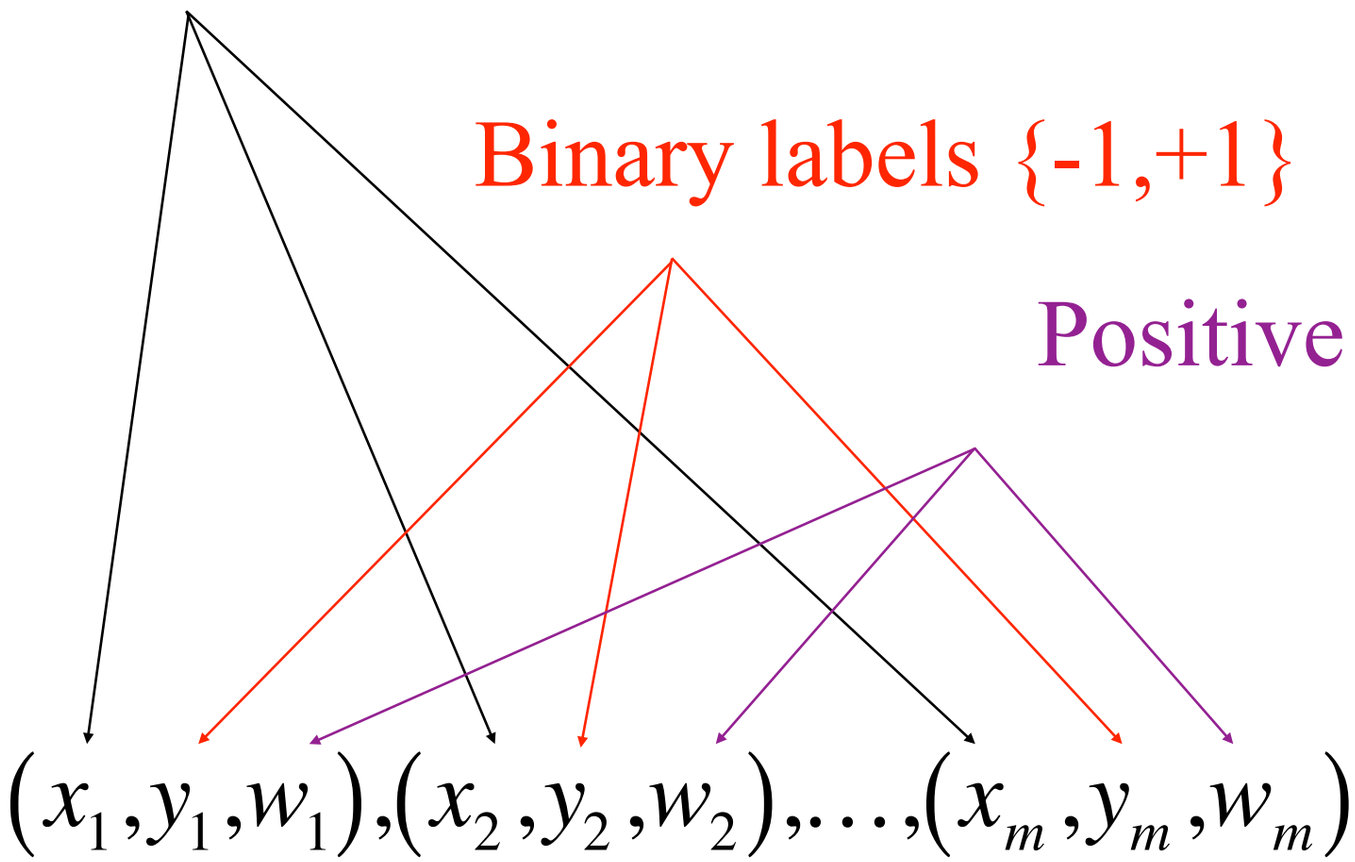
Adaboost

A weighted training set

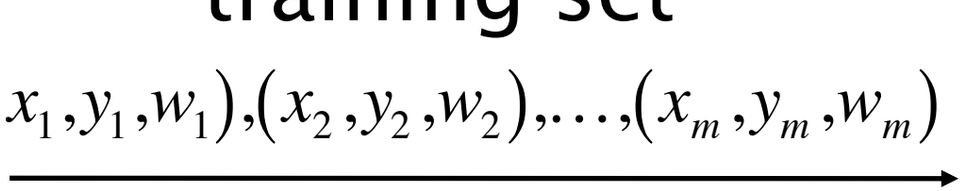
Feature vectors



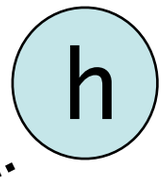
A weak learner

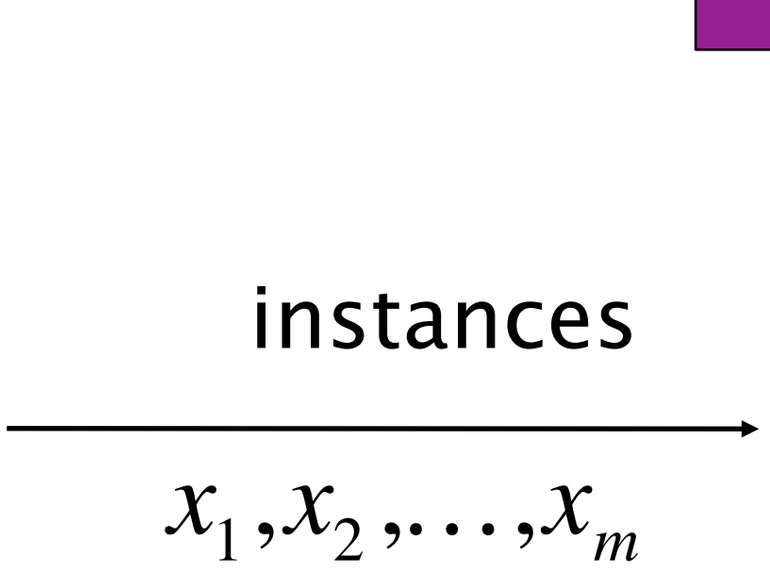
Weighted training set

$(x_1, y_1, w_1), (x_2, y_2, w_2), \ldots, (x_m, y_m, w_m)$



A weak rule





instances

$x_1, x_2, \ldots, x_m$

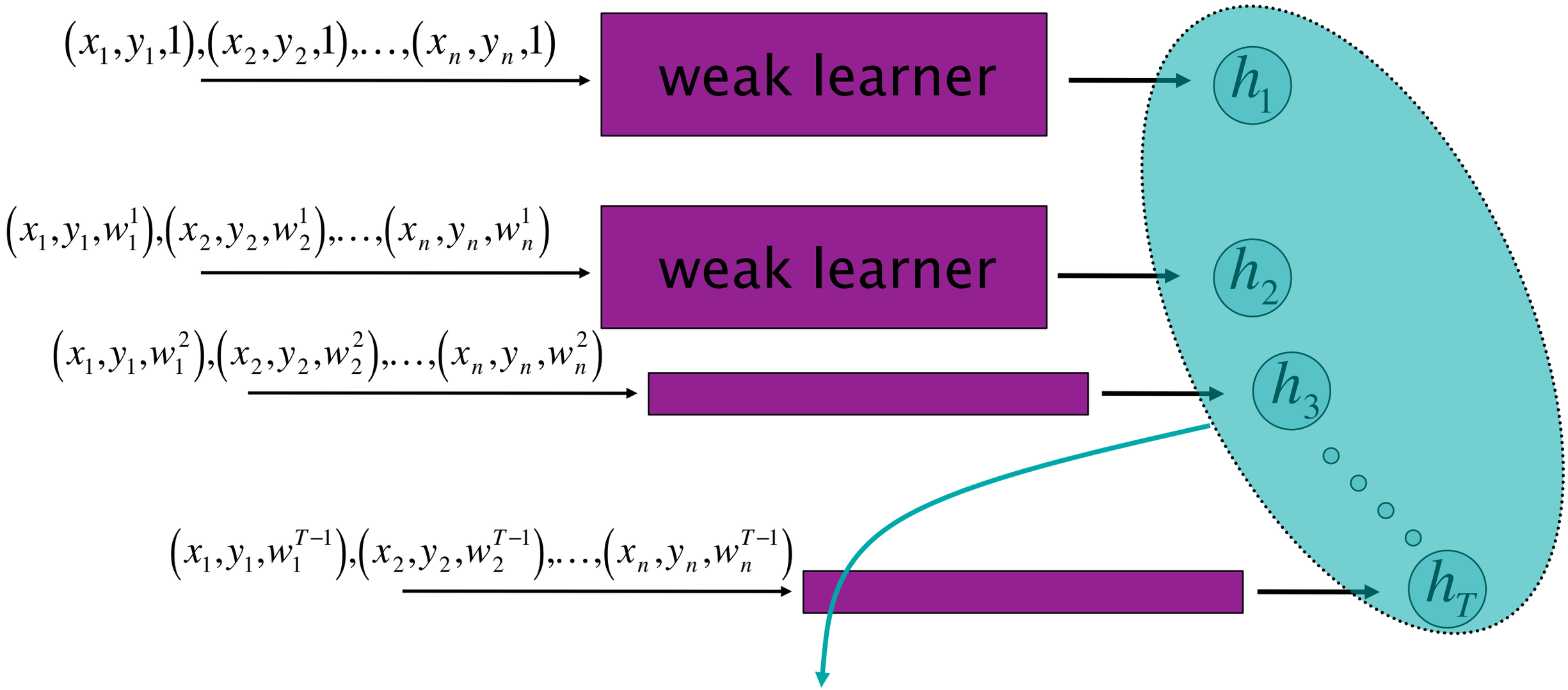
predictions

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m; \quad \hat{y}_i \in \{0,1\}$

The weak requirement:

$$\left| \frac{\sum_{i=1}^{m} y_i \hat{y}_i w_i}{\sum_{i=1}^{m} w_i} \right| > \gamma > 0$$

The boosting process



$$F_T(x) = \alpha_1 h_1(x) + \alpha_2 h_2(x) + \ldots + \alpha_T h_T(x)$$

Adaboost

Freund, Schapire 1997

$$F_0(x) \equiv 0$$

$$\text{for } t = 1..T$$

$$w_i^t = \exp\left(-y_i F_{t-1}(x_i)\right)$$

$$\text{Get } h_t \text{ from } weak-learner$$

$$\alpha_t = \frac{1}{2}\ln\left(\sum_{i:h_t(x_i)=1, y_i=1} w_i^t \Big/ \sum_{i:h_t(x_i)=1, y_i=-1} w_i^t\right)$$

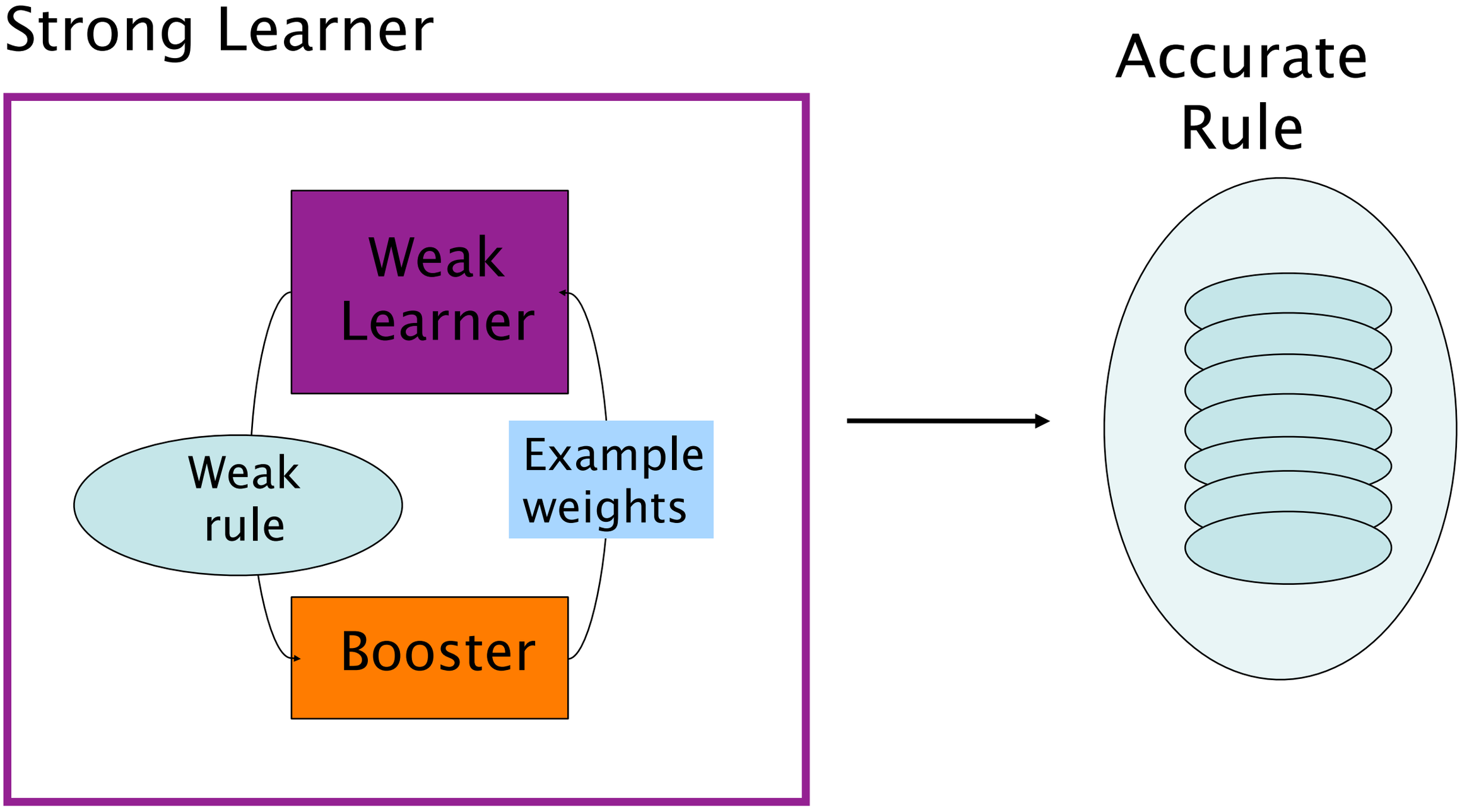
$$F_{t+1} = F_t + \alpha_t h_t$$

Main property of Adaboost

If advantages of weak rules over random guessing are: $\gamma_1,\gamma_2,..,\gamma_T$ then training error of final rule is at most

$$\hat{\varepsilon}(f_T) \le \exp\left(-\sum_{t=1}^{T}\gamma_t^2\right)$$

Boosting block diagram

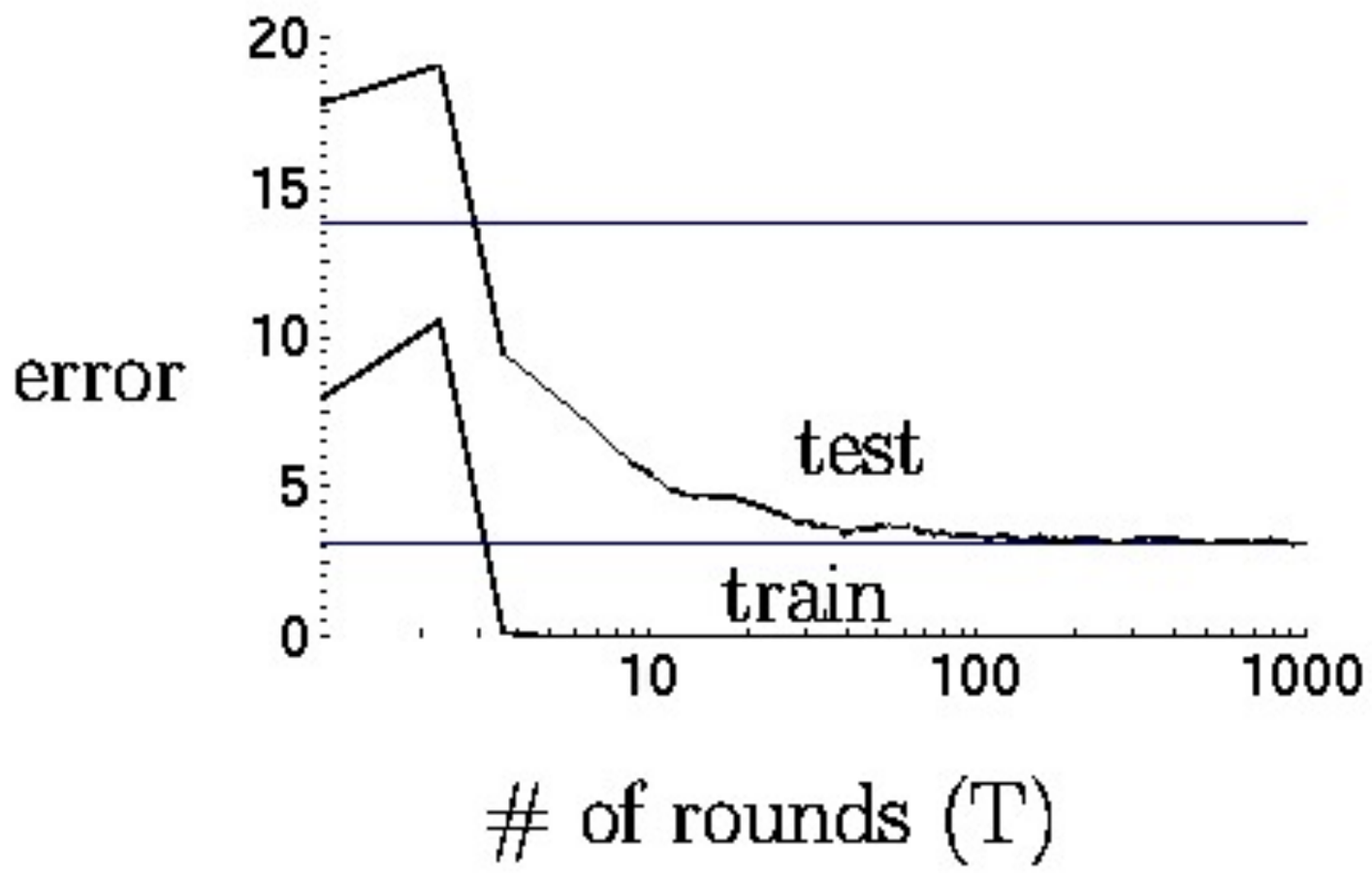


Boosting the margins,
Over-fitting,
Bias,
Variance

and all that Jazz

A very curious phenomenon

Boosting decision trees



Using <10,000 training examples we fit >2,000,000 parameters

Large margins

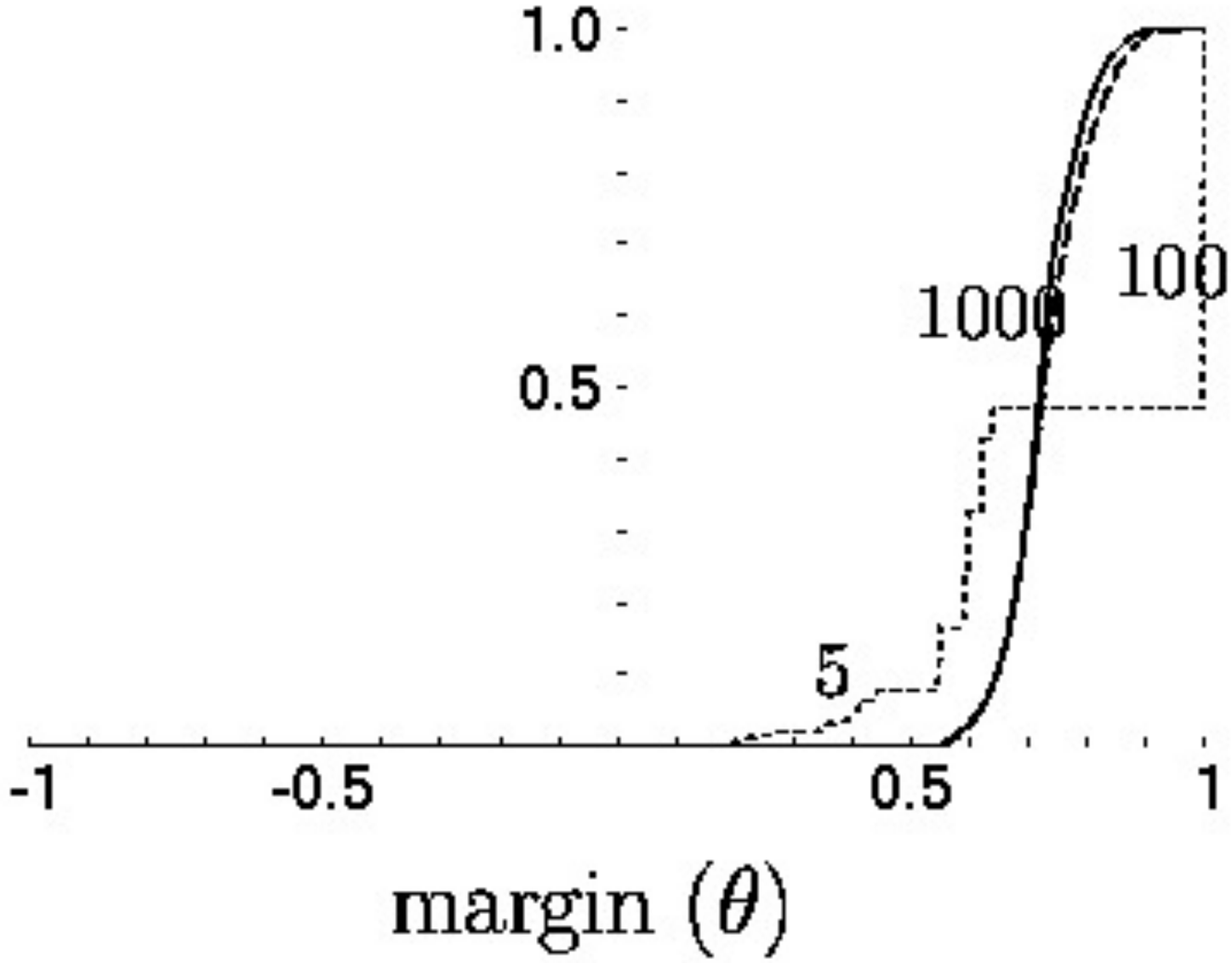
$$\text{margin}_{F_T}(x,y) \doteq y \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} |\alpha_t|} = y \frac{F_T(x)}{\|\vec{\alpha}\|_1}$$

$$\text{margin}_{F_T}(x,y) > 0 \quad \Leftrightarrow \quad f_T(x) = y$$

Thesis:
large margins => reliable predictions

Very similar to SVM.

Experimental Evidence



Prediction uncertainty

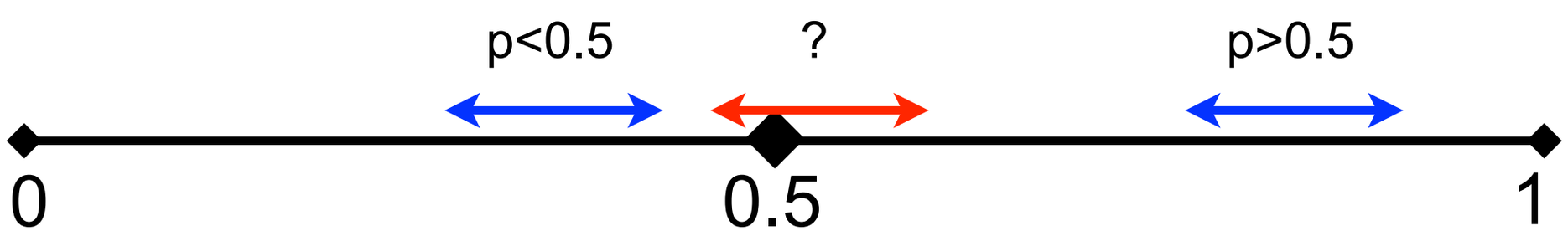
versus

Training uncertainty

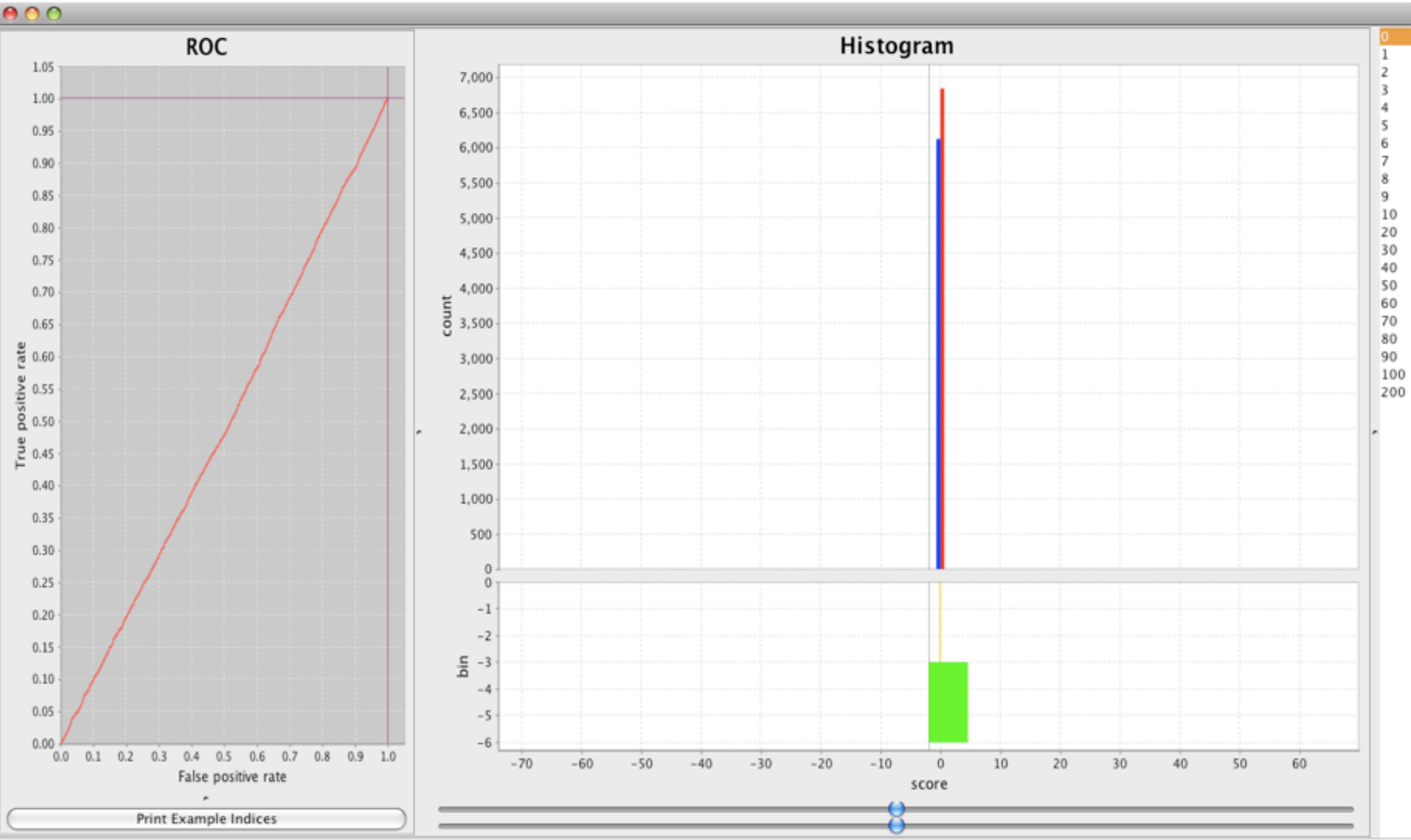
- Prediction uncertainty:
  P(label | Instance)
- Training uncertainty:
  Distance btwn estimate of P and true P.
- Margins measure training uncertainty, NOT prediction uncertainty.

Does Boosting reduce Bias or Variance?

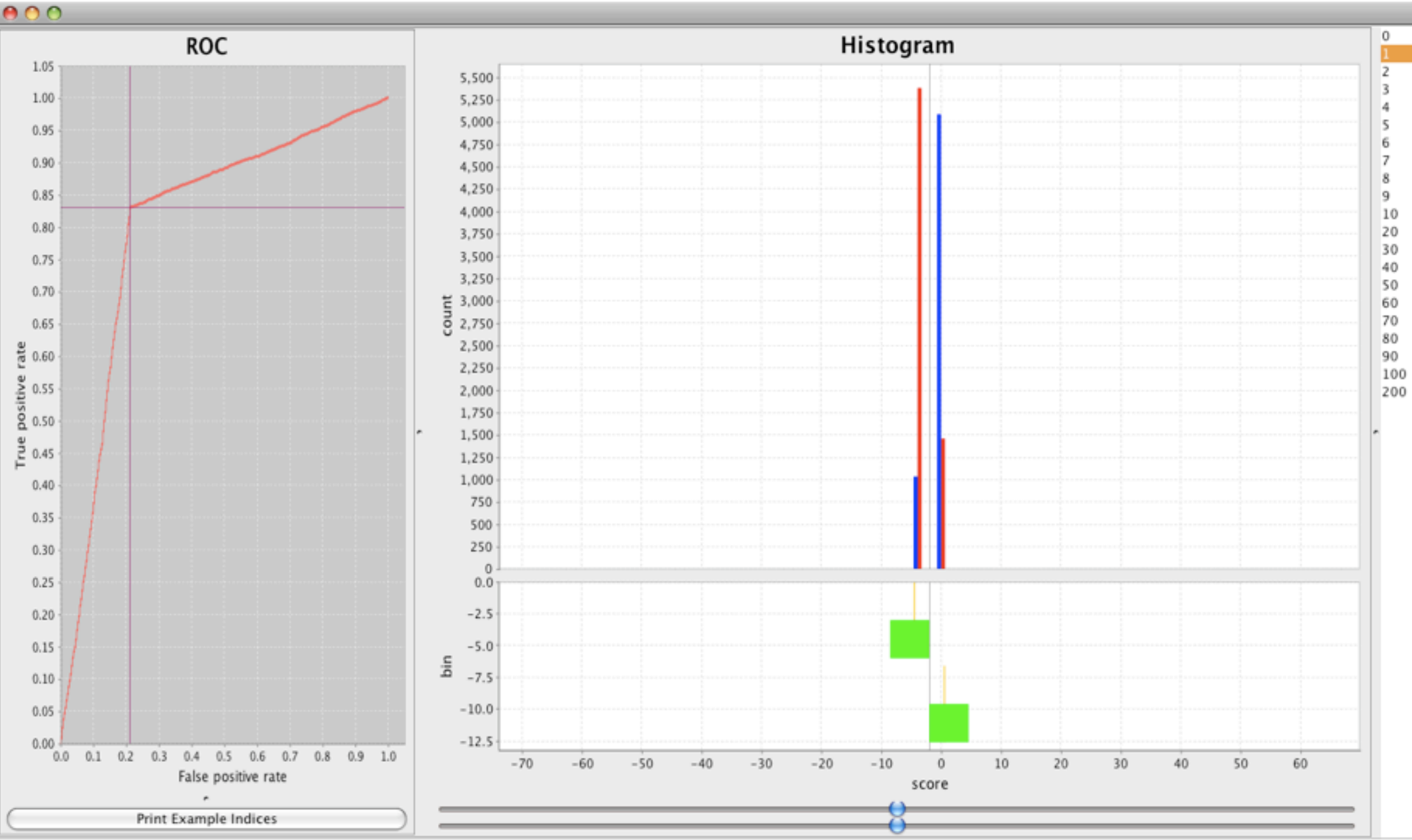
- **Ill-defined question**: Bias and Variance defined for regression, not classification.
- For classification, required accuracy of conditional probability estimate depends on the distance from 1/2



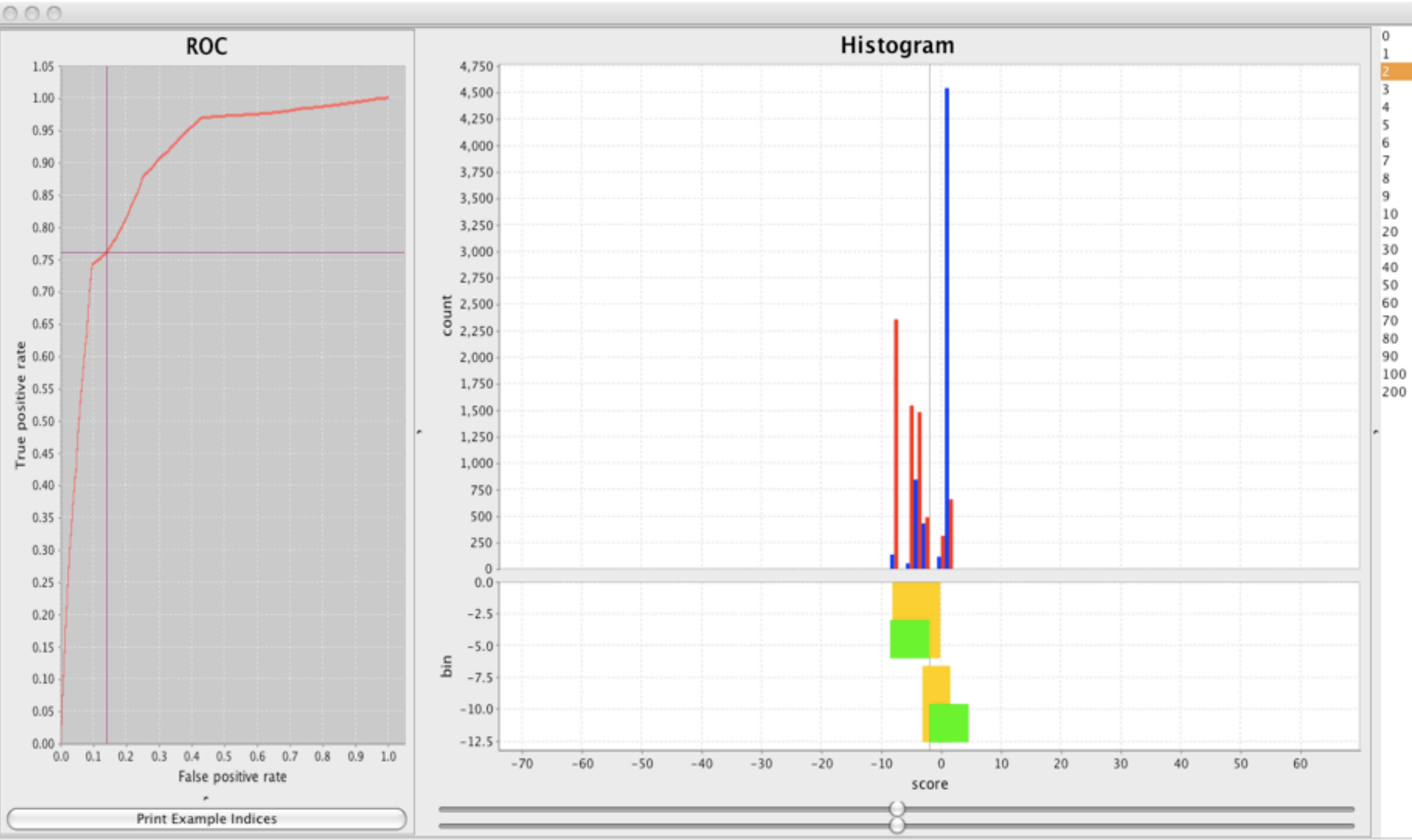
Iteration 0


ROC
True positive rate
False positive rate
Print Example Indices
Histogram
count
bin
score

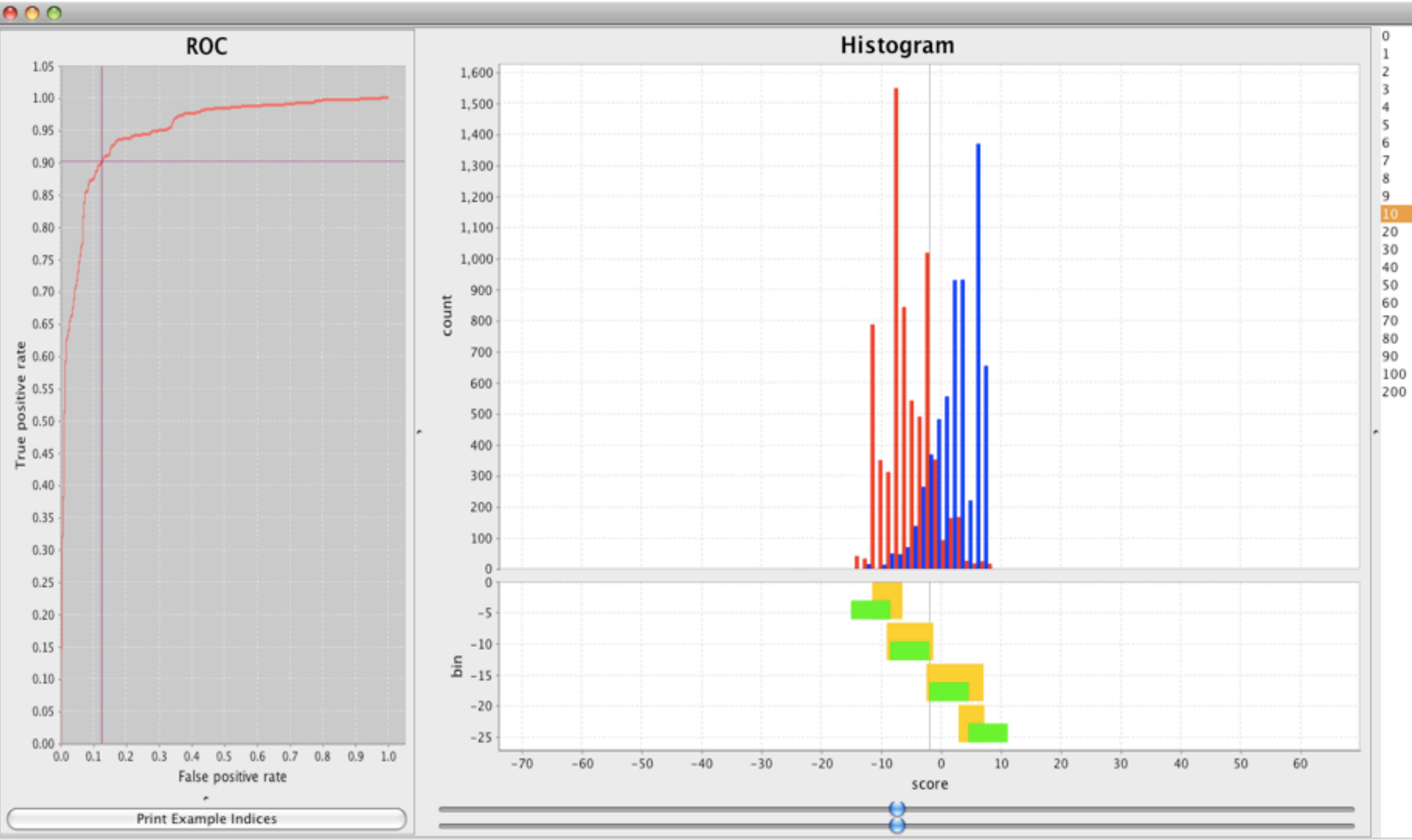
Iteration 1


ROC
True positive rate
False positive rate
Print Example Indices
Histogram
count
bin
score

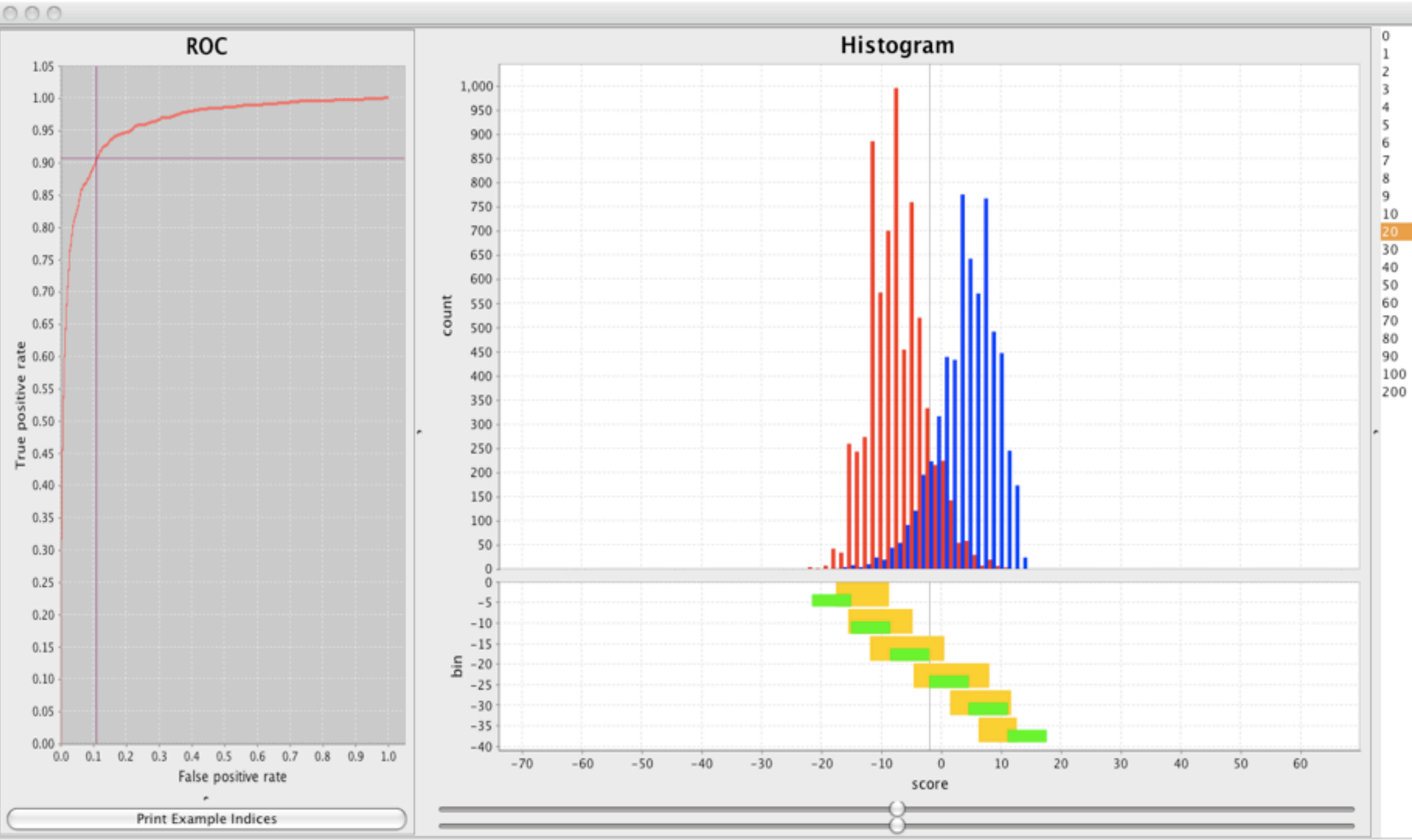
Iteration 2


ROC
True positive rate
False positive rate
Print Example Indices
Histogram
count
bin
score

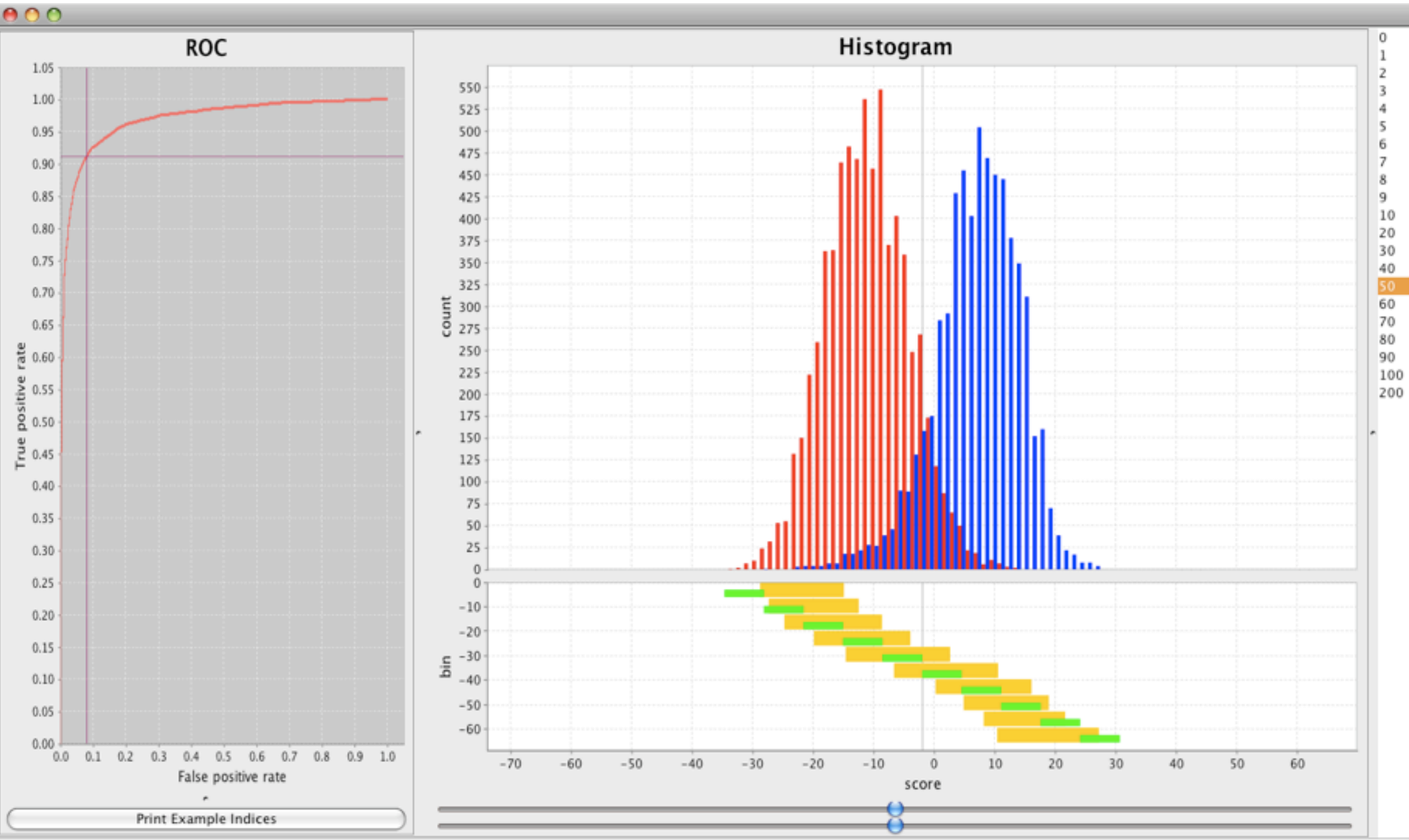
Iteration 10



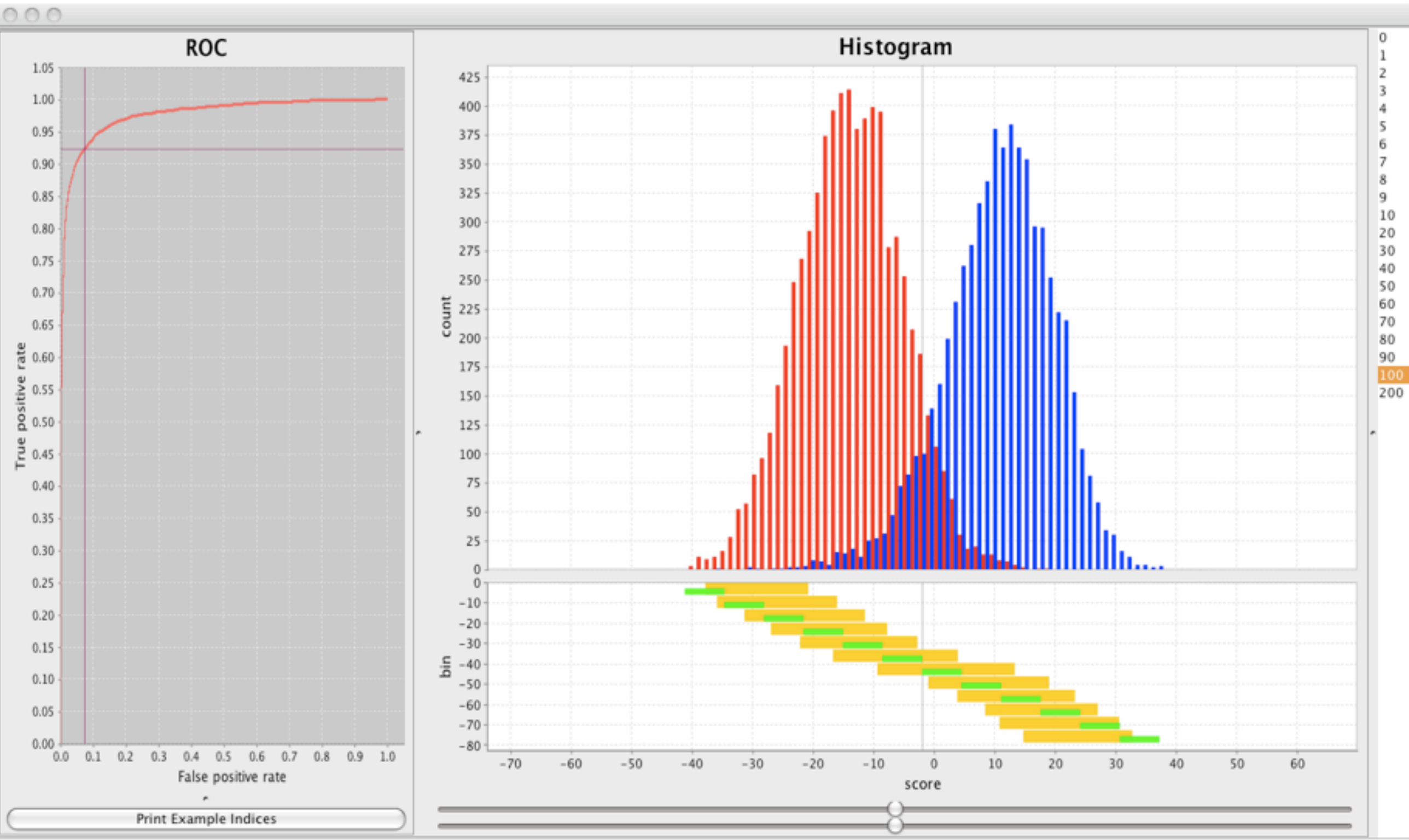
Iteration 20



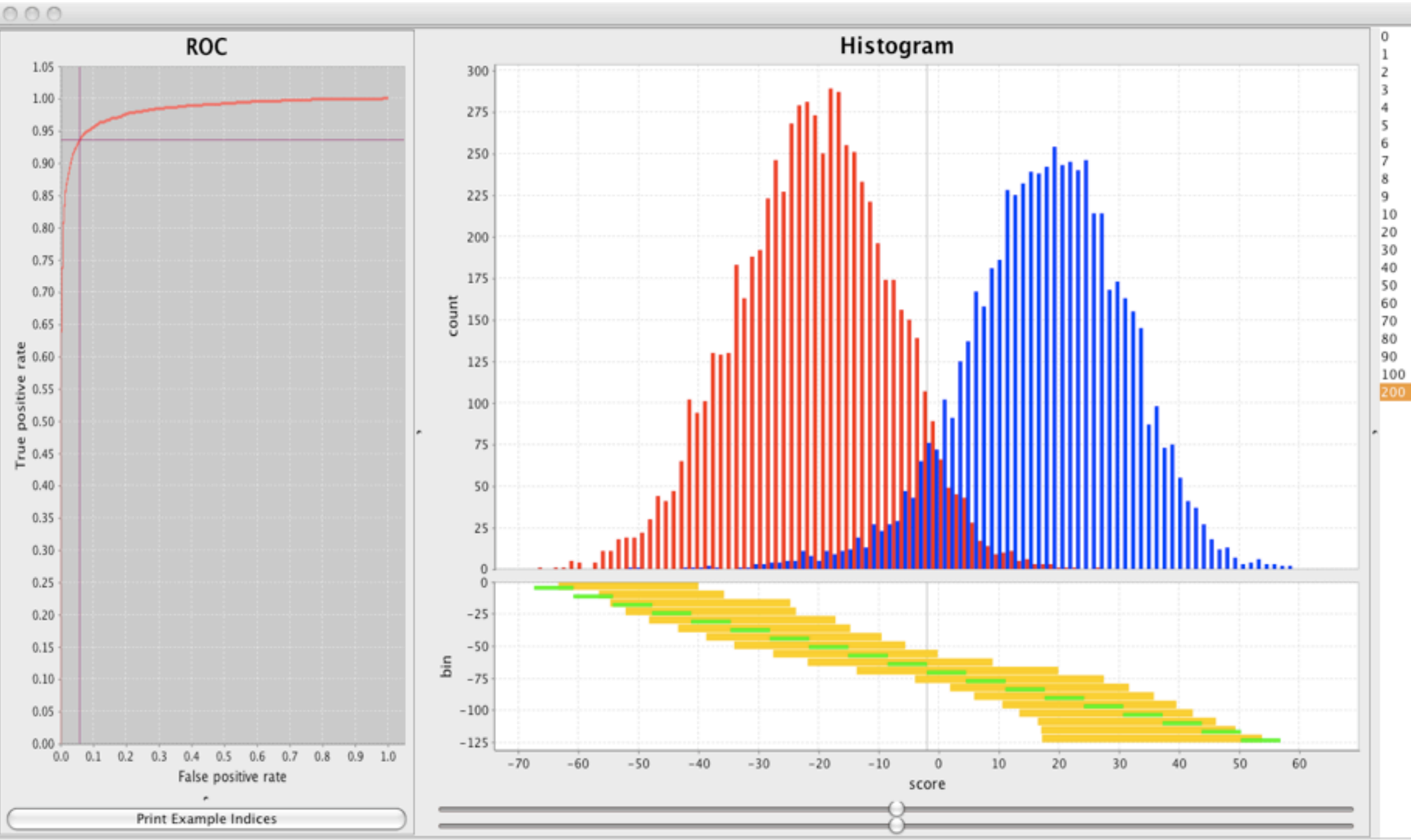
Iteration 50



Iteration 100



Iteration 200



scores after retraining