

True positive rate (Sensitivity)

$$\text{true positive rate} = \frac{\# \text{ of true positives}}{\# \text{ of known positives}}$$

(Proportion of actual positives that are correctly identified)

True negative rate (Specificity)

$$\text{true negative rate} = \frac{\# \text{ of true negatives}}{\# \text{ of known negatives}}$$

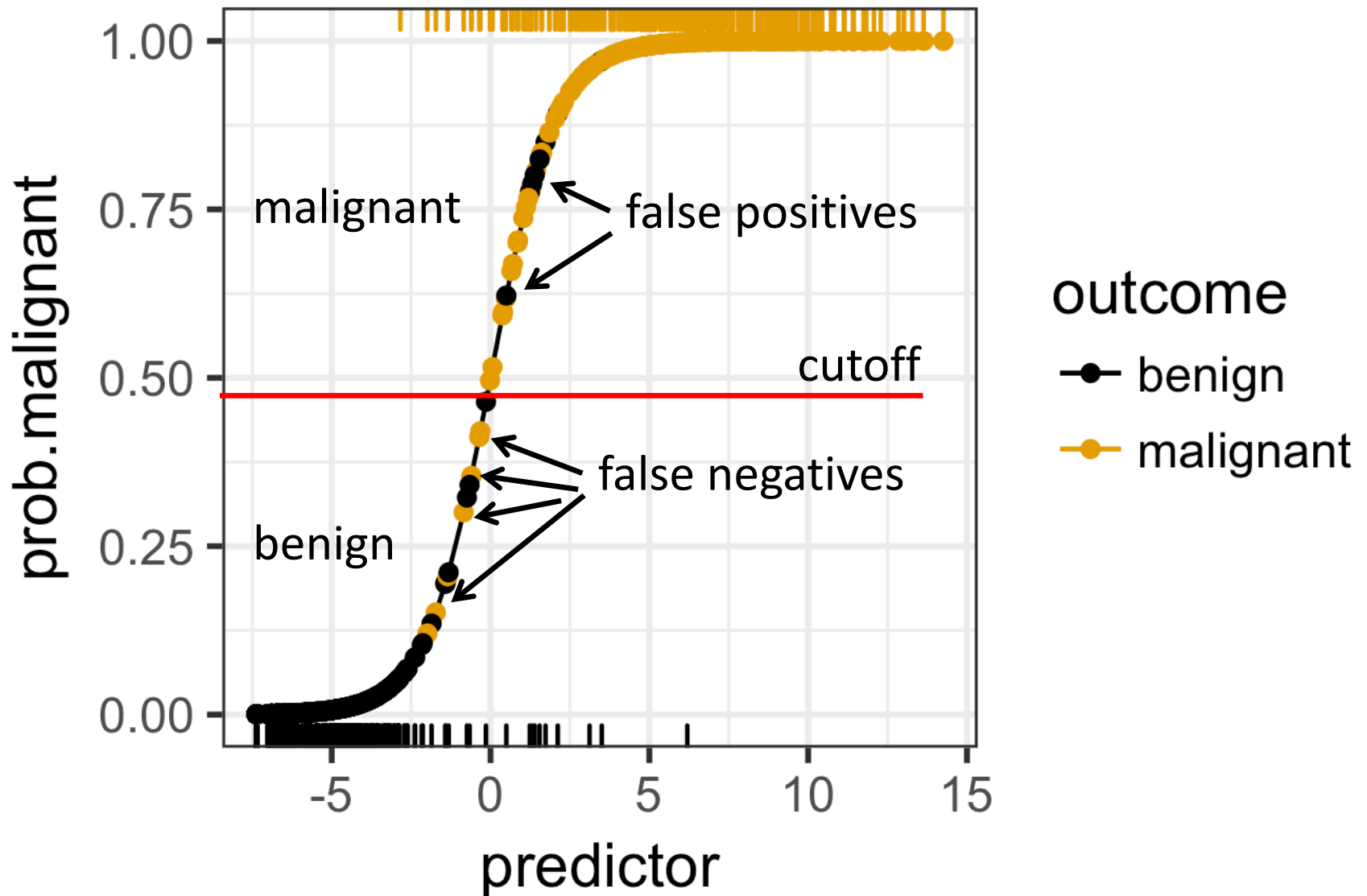
(Proportion of actual negatives that are correctly identified)

False positive rate (1 – Specificity)

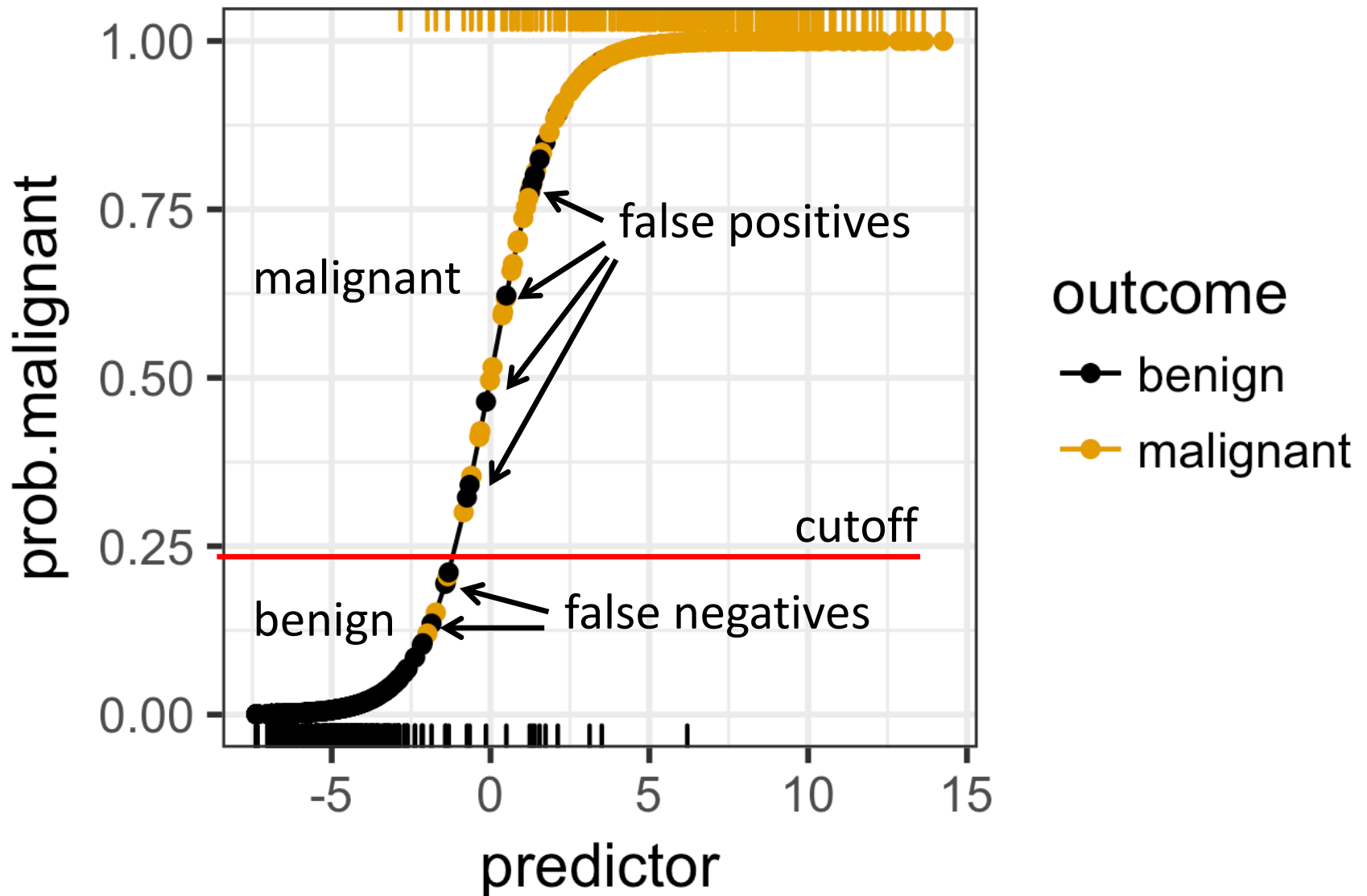
$$\text{false positive rate} = \frac{\# \text{ of false positives}}{\# \text{ of known negatives}}$$

(Proportion of actual negatives that are **incorrectly** identified)

Sensitivity and specificity depend on a chosen cutoff

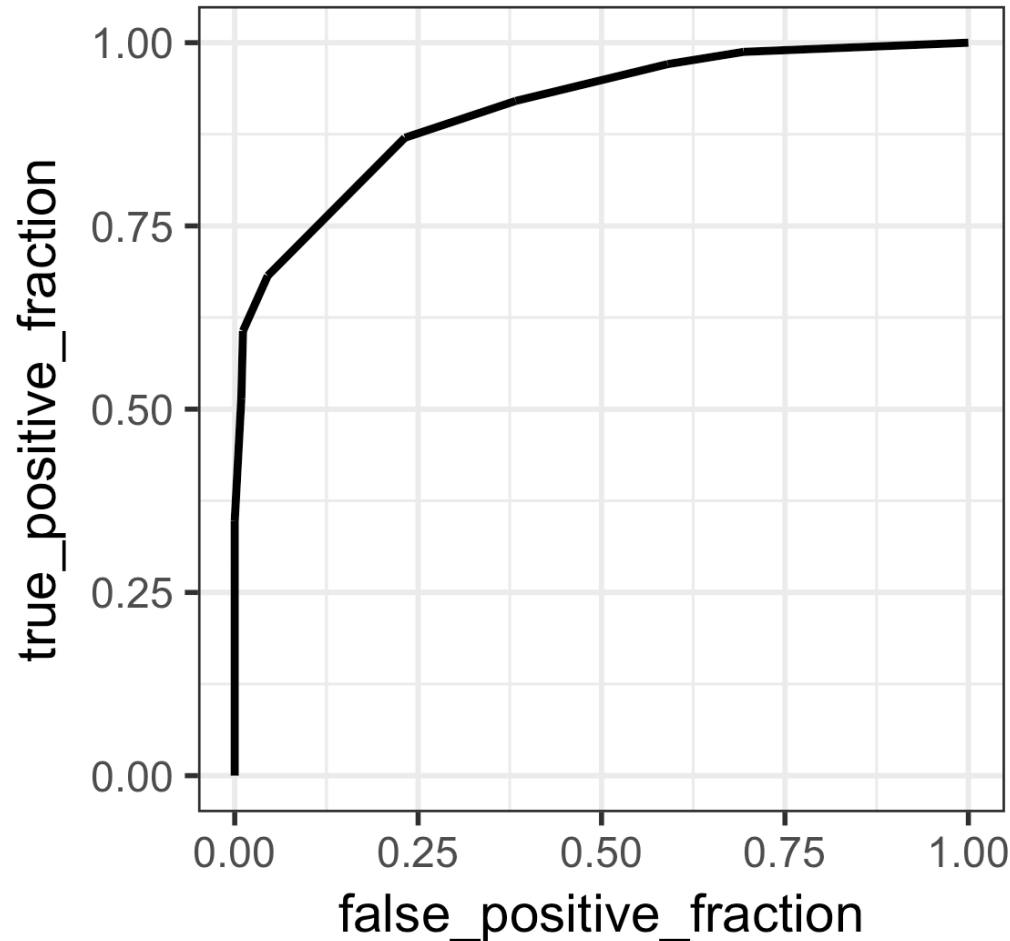


Sensitivity and specificity depend on a chosen cutoff

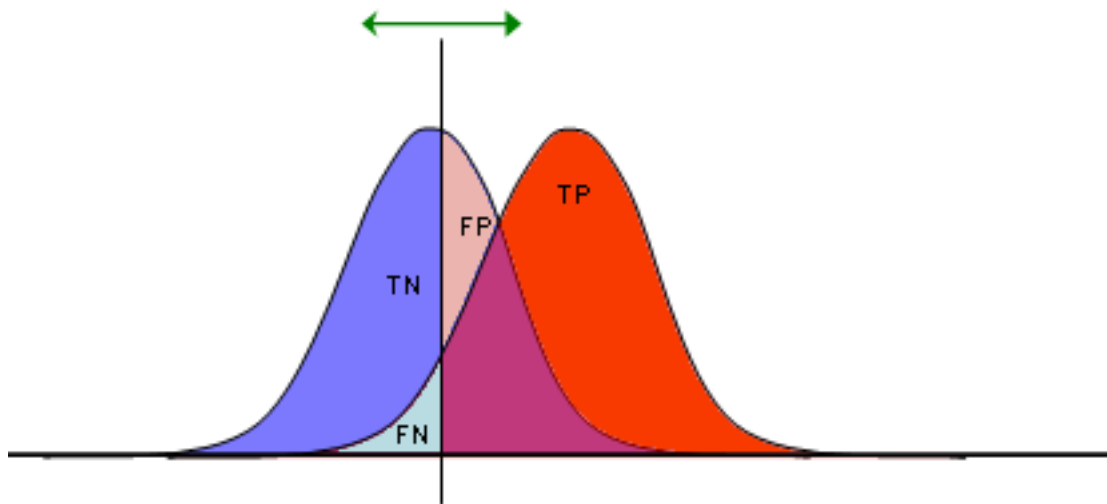


Do Part 1 of the worksheet now

We usually plot the true pos. rate vs. the false pos. rate for all possible cutoffs



ROC curve
Receiver
Operating
Characteristic
curve



TP	FP
FN	TN
1	1

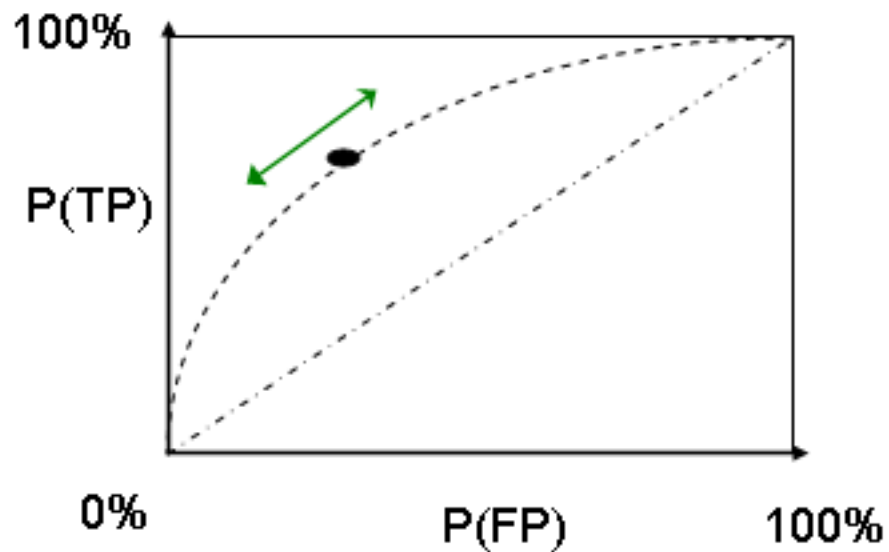
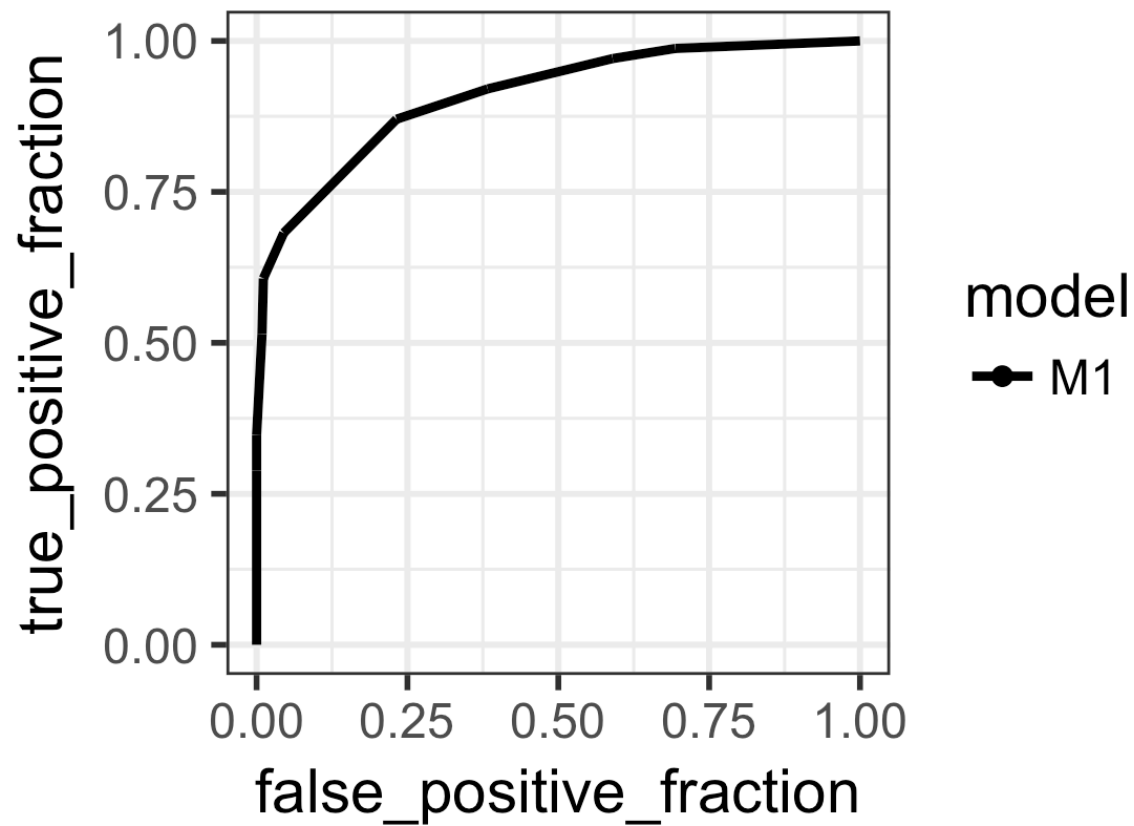


Image from: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

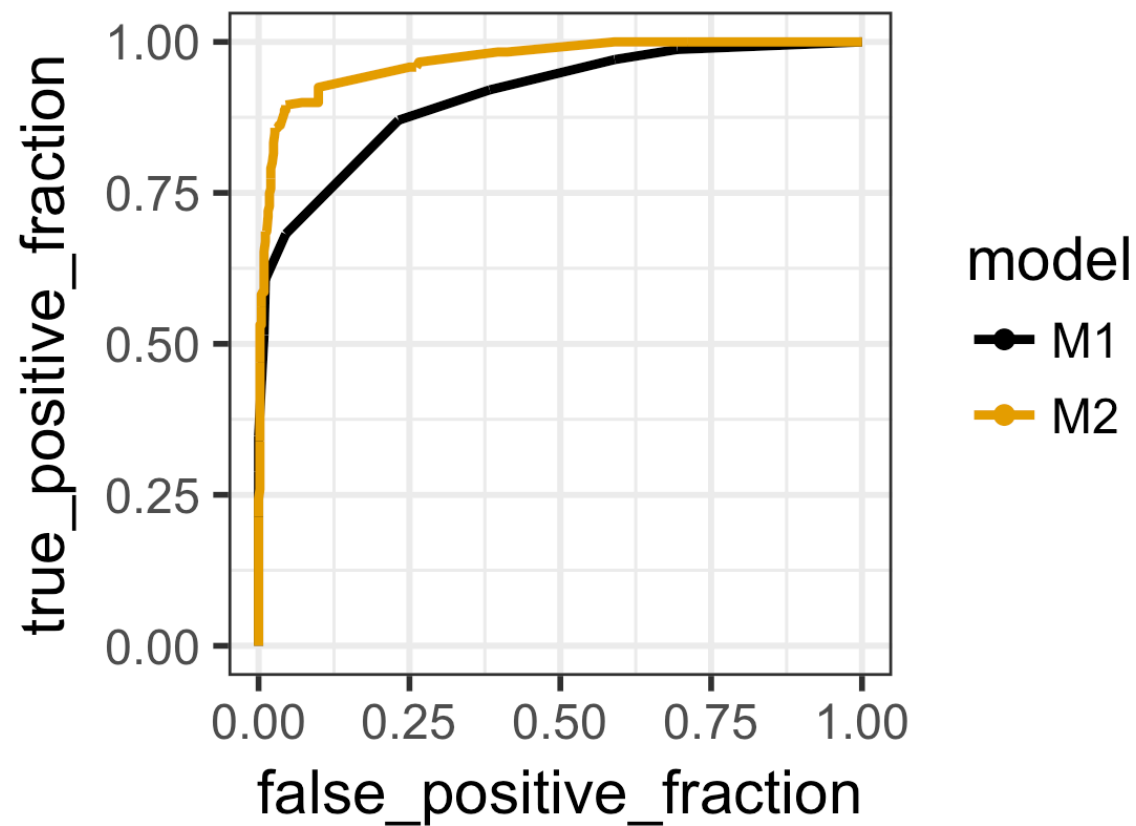
The area under the curve tells us how good a model's predictions are



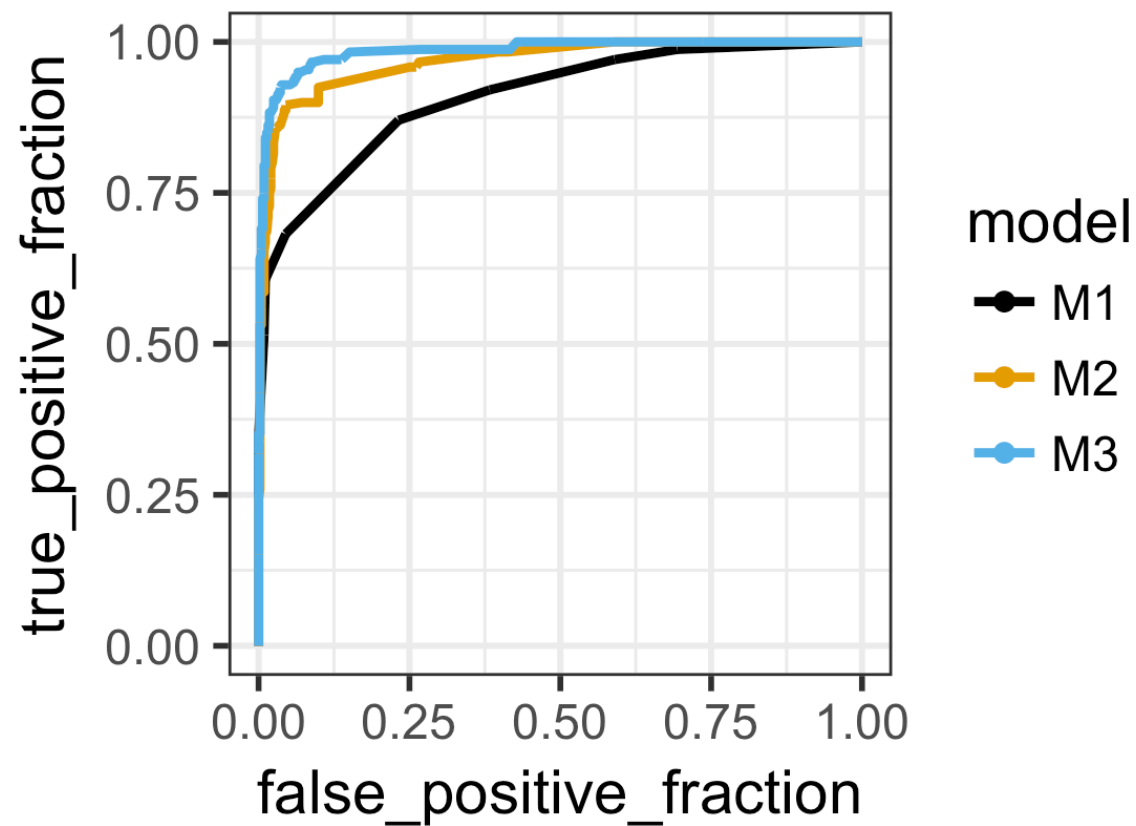
Let's look at the performance of several different models for the biopsy data set



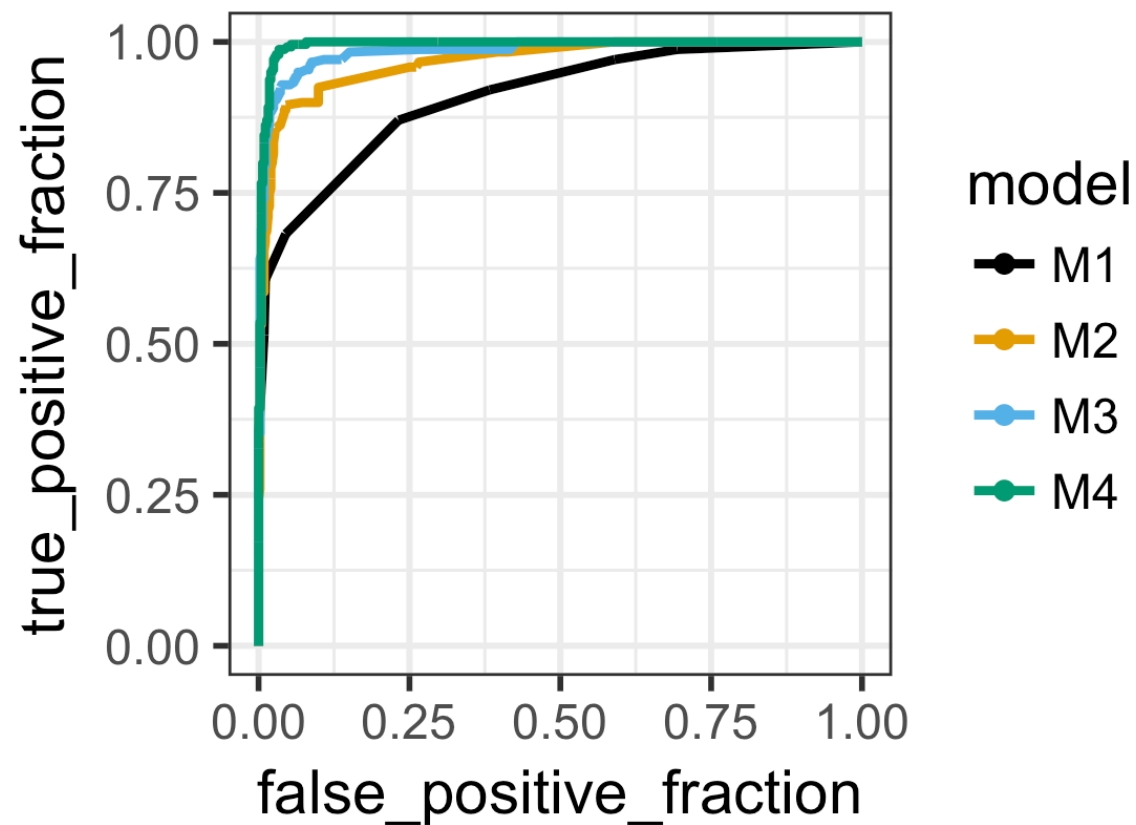
Predictor	M1
clump_thickness	✓
normal_nucleoli	
marg_adhesion	
bare_nuclei	
uniform_cell_shape	
bland_chromatin	



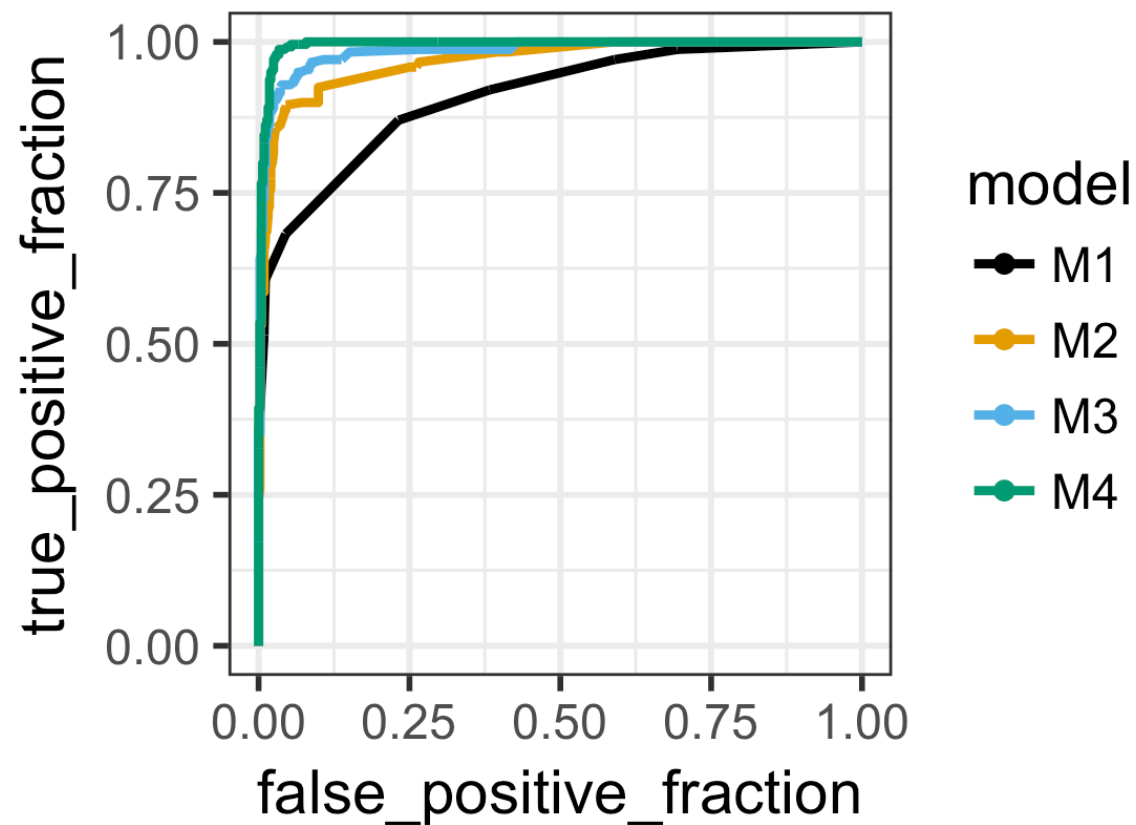
Predictor	M1	M2
clump_thickness	✓	✓
normal_nucleoli		✓
marg_adhesion		
bare_nuclei		
uniform_cell_shape		
bland_chromatin		



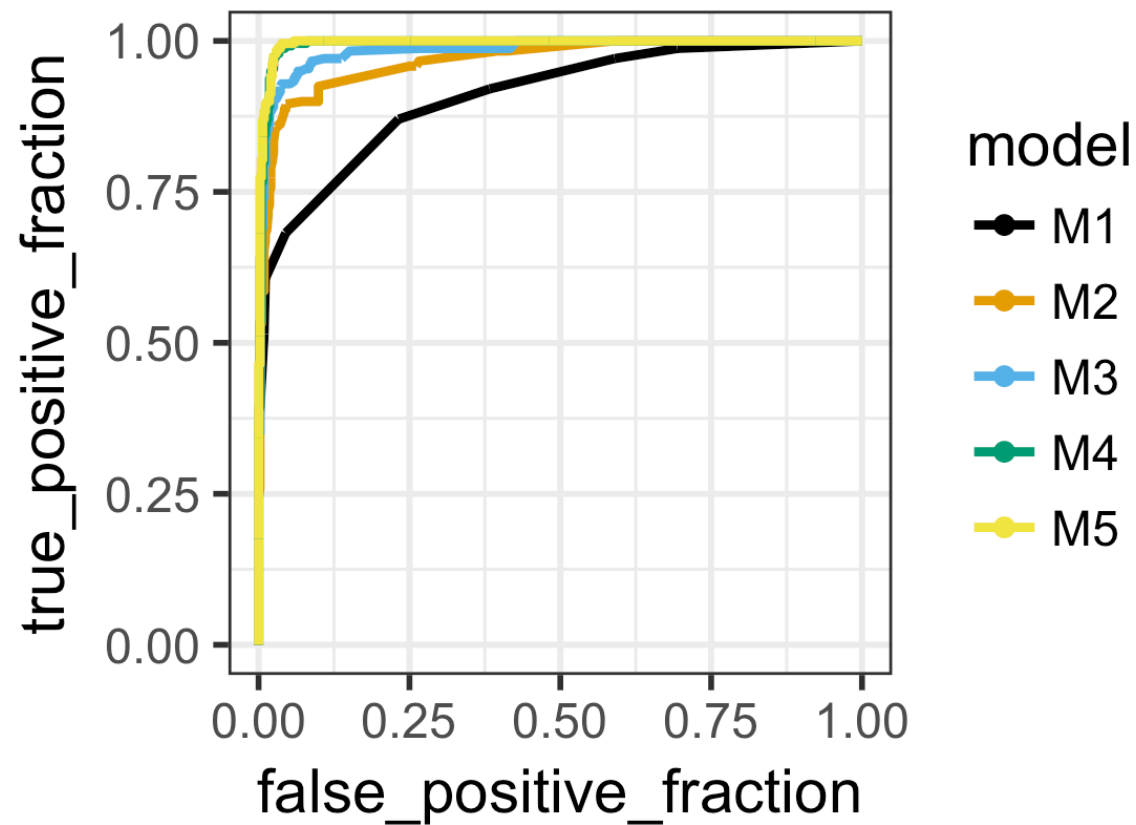
Predictor	M1	M2	M3
clump_thickness	✓	✓	✓
normal_nucleoli		✓	✓
marg_adhesion			✓
bare_nuclei			
uniform_cell_shape			
bland_chromatin			



Predictor	M1	M2	M3	M4
clump_thickness	✓	✓	✓	✓
normal_nucleoli		✓	✓	✓
marg_adhesion			✓	✓
bare_nuclei				✓
uniform_cell_shape				
bland_chromatin				

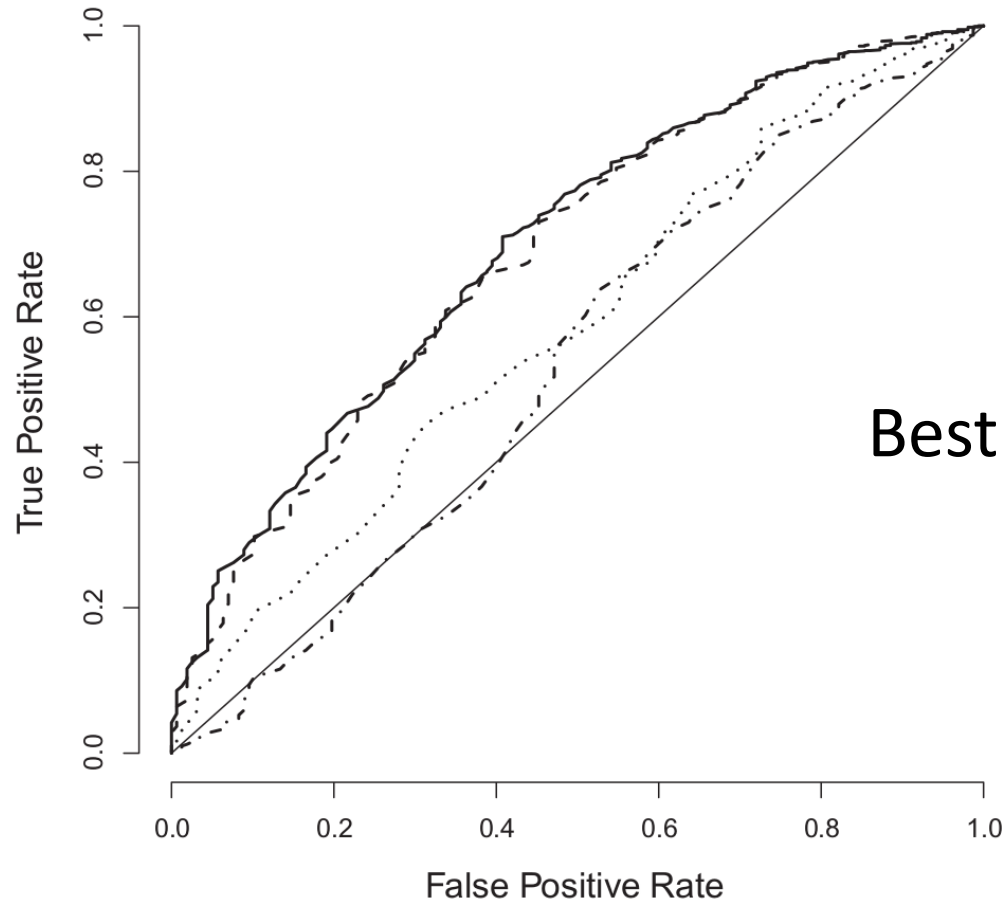


Predictor	M1	M2	M3	M4	M5
clump_thickness	✓	✓	✓	✓	✓
normal_nucleoli		✓	✓	✓	✓
marg_adhesion			✓	✓	✓
bare_nuclei				✓	✓
uniform_cell_shape					✓
bland_chromatin					✓



Model	Area Under Curve (AUC)
M1	0.909
M2	0.968
M3	0.985
M4	0.995
M5	0.996

Things usually look much worse in real life



Best AUC (solid line): 0.70

Calculating ROC curves in R

Using `geom_roc()` from the `plotROC` package

Using `geom_roc()` from the `plotROC` package

```
# fit a logistic regression model  
glm.out <- glm(outcome ~ clump_thickness,  
               data=biopsy, family = binomial)
```

Using `geom_roc()` from the `plotROC` package

```
# fit a logistic regression model
glm.out <- glm(outcome ~ clump_thickness,
               data=biopsy, family = binomial)

# calculate ROC curve
df <- data.frame(probabilities = predict(glm.out, biopsy),
                 known_truth = biopsy$outcome,
                 model = 'M1')
```

Using `geom_roc()` from the `plotROC` package

```
# fit a logistic regression model
glm.out <- glm(outcome ~ clump_thickness,
               data=biopsy, family=binomial)

# prepare data for ROC plotting
df <- data.frame(probabilities = predict(glm.out, biopsy),
                 known_truth = biopsy$outcome,
                 model = 'M1')

# the aesthetic names are not the most intuitive
# `d` (disease) holds the known truth
# `m` (marker) holds the predictor values
p <- ggplot(df, aes(d = known_truth, m = predictor)) +
  geom_roc(n.cuts = 0) + coord_fixed()

p # make plot
```

Calculating the area under the curve (AUC)

```
# the function calc_auc needs to be called on a plot object  
# that uses geom_roc():  
calc_auc(p)
```

```
#   PANEL group      AUC  
# 1      1      -1 0.908878  
# Warning message:  
# In verify_d(data$d) :  
#   D not labeled 0/1, assuming benign = 0 and malignant = 1!
```

Do Part 2 of the worksheet now