Machine Learning in R

Lecture 2 — Classification

Jean Feng – Epidemiology and Biostatistics

Any questions from last lecture?

1. What is Machine Learning?
2. Supervised vs Unsupervised Learning
3. Supervised learning:
    1. Notation
    2. Evaluating models using a loss function
    3. Parametric and non-parametric models
    4. Model selection
    5. Bias-Variance Trade-off
    6. Multivariate models

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. K-nearest neighbors
    6. Receiver Operating Characteristic (ROC)

Today's lecture

1. **Linear regression as an optimization procedure**
2. Classification
    1. Loss functions
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. K-nearest neighbors
    6. Receiver Operating Characteristic (ROC)

Recall: Supervised Learning

- **Response**: what we want to predict, also known as the outcome, target, or dependent variable

$$Y$$

- **Predictors**: inputs, regressors, covariates, features, or variables

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

- **Dataset**:

$$\{(x_i, y_i) : i = 1,...,n\}$$

- **Model**: A function $f(x)$ that predicts Y at point $X = x$

Linear regression

The best linear model is defined as

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \mathbb{E}\left[\left(Y - \underbrace{(X^\top \beta + \beta_0)}_{\text{Prediction } f(X)}\right)^2\right]$$

Ordinary least squares fits a model by solving the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (x_i^\top \beta + \beta_0)\right)^2$$

More generally…

If we define ℓ as the **squared error loss**, then

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (x_i^\top \beta + \beta_0) \right)^2$$

is equivalent to

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right).$$

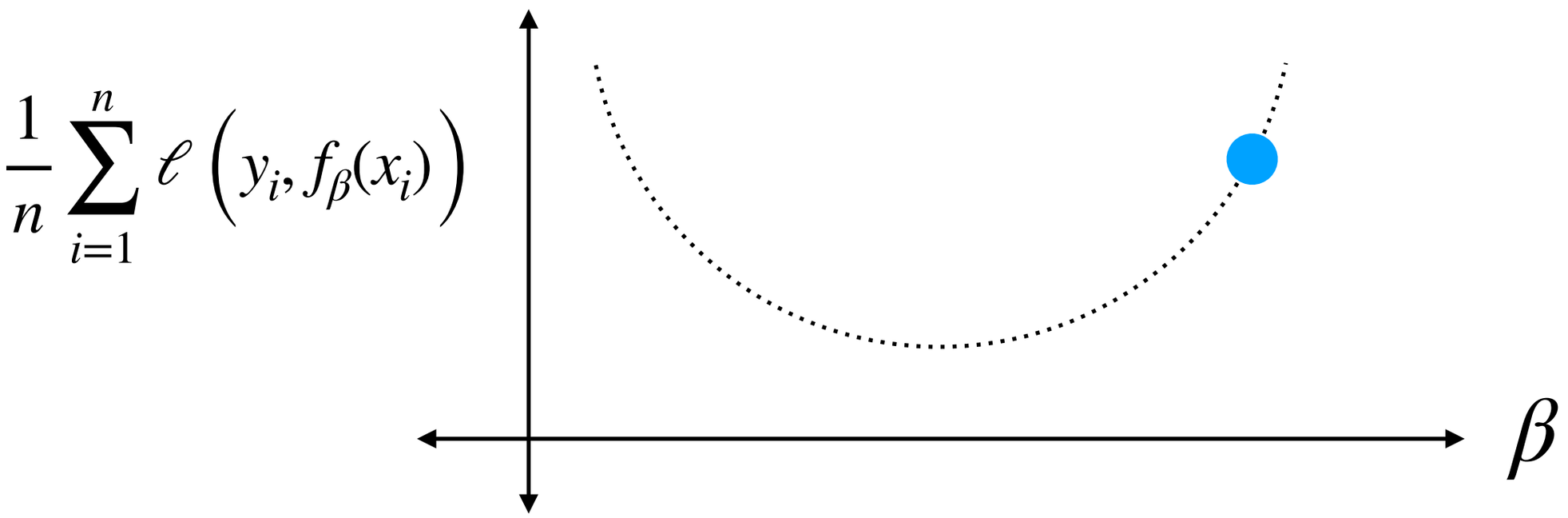
Many machine learning algorithms can be written in a similar way, where it is defined as the minimizer of the average loss.

Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:



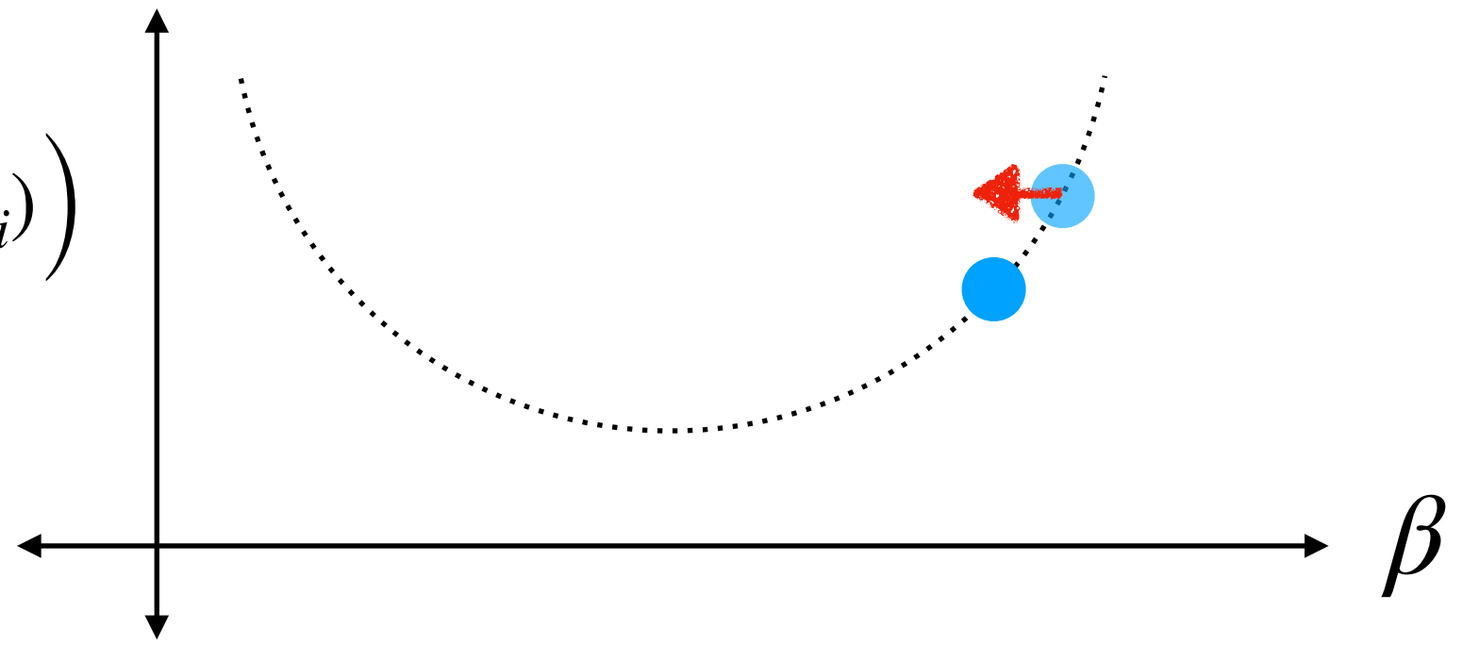
Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:

$$\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)$$



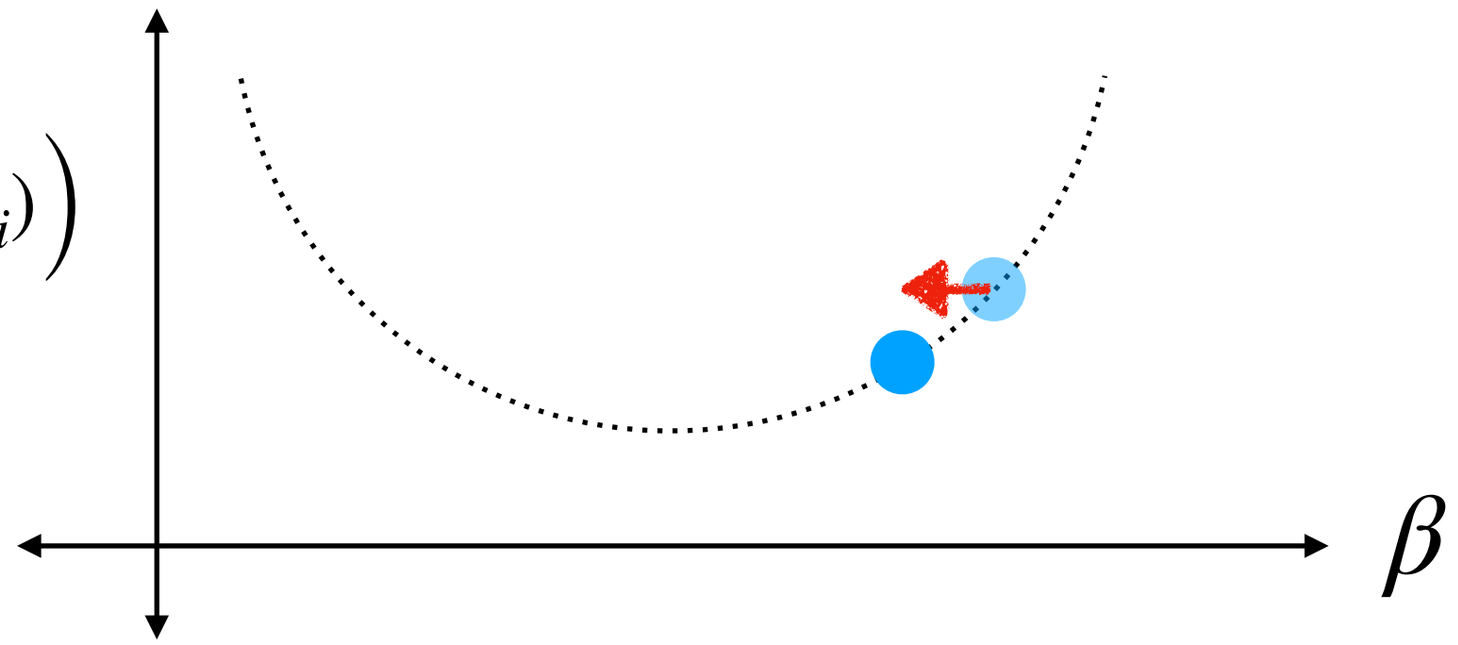
Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:

$$\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)$$



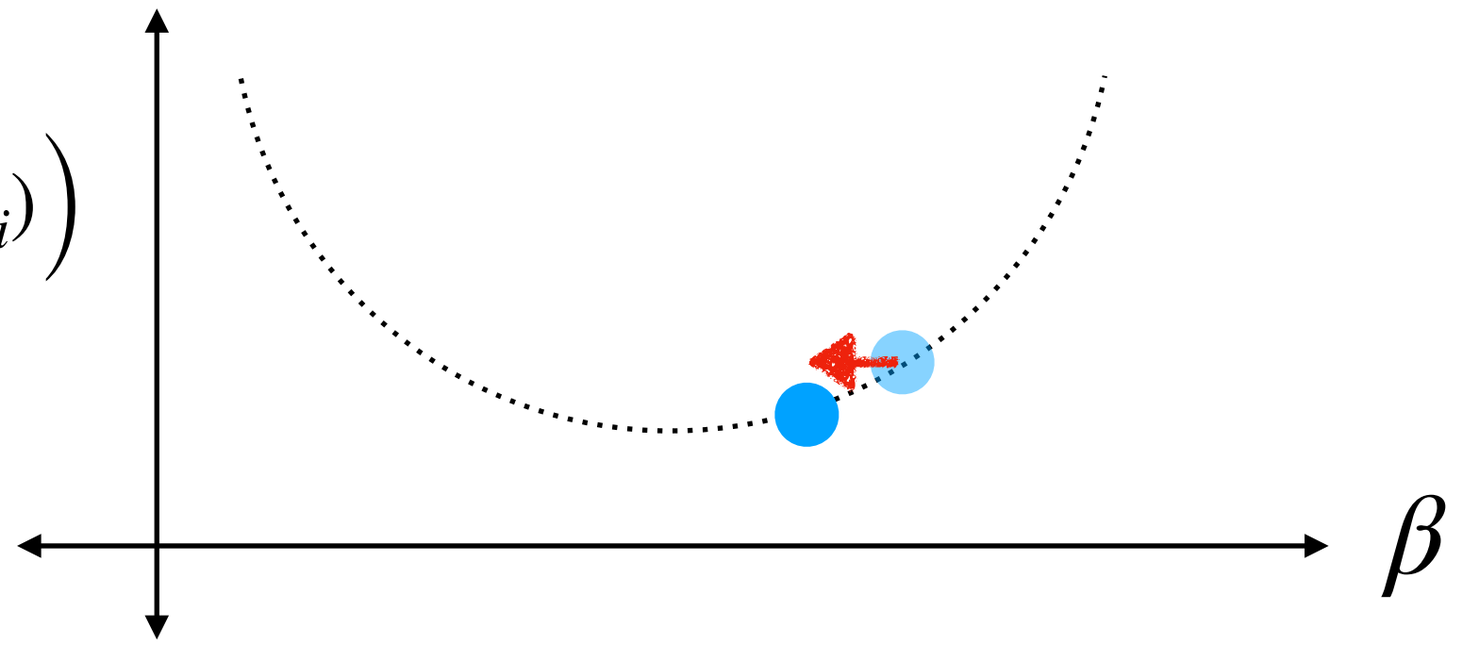
Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:

$$\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)$$



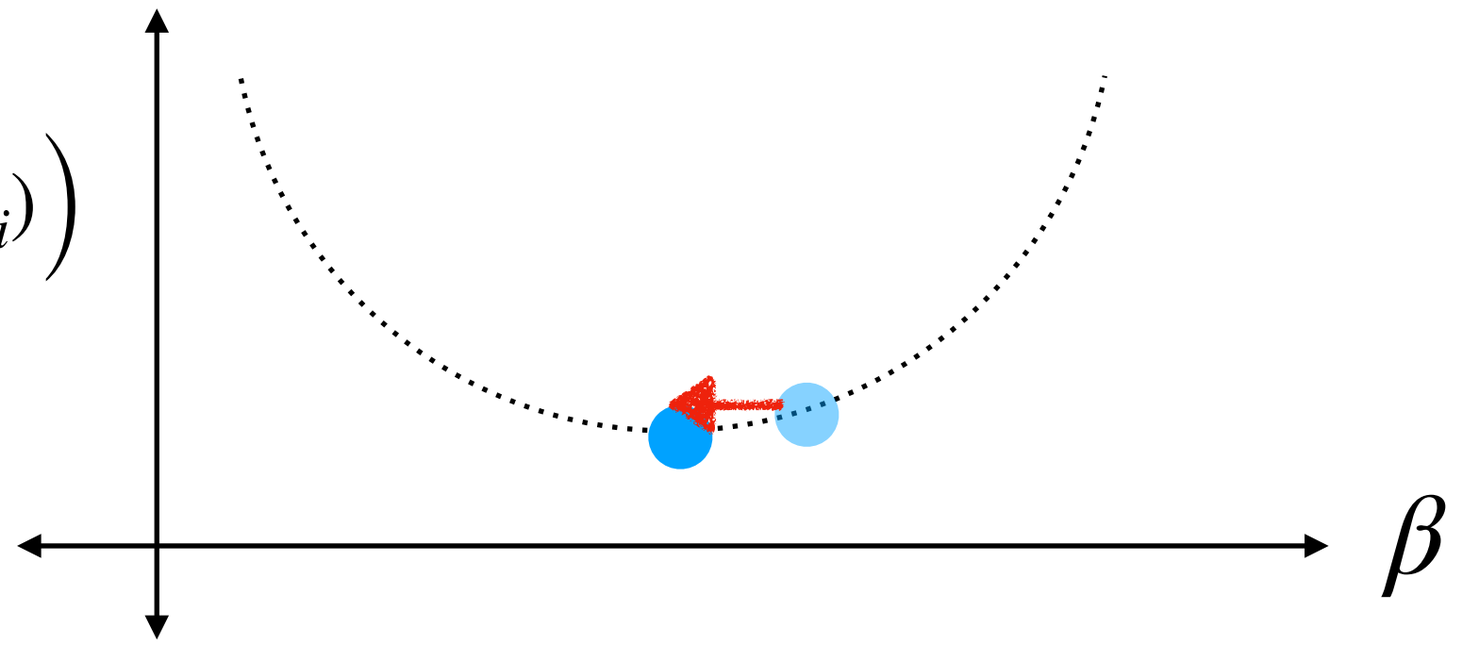
Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:

$$\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)$$



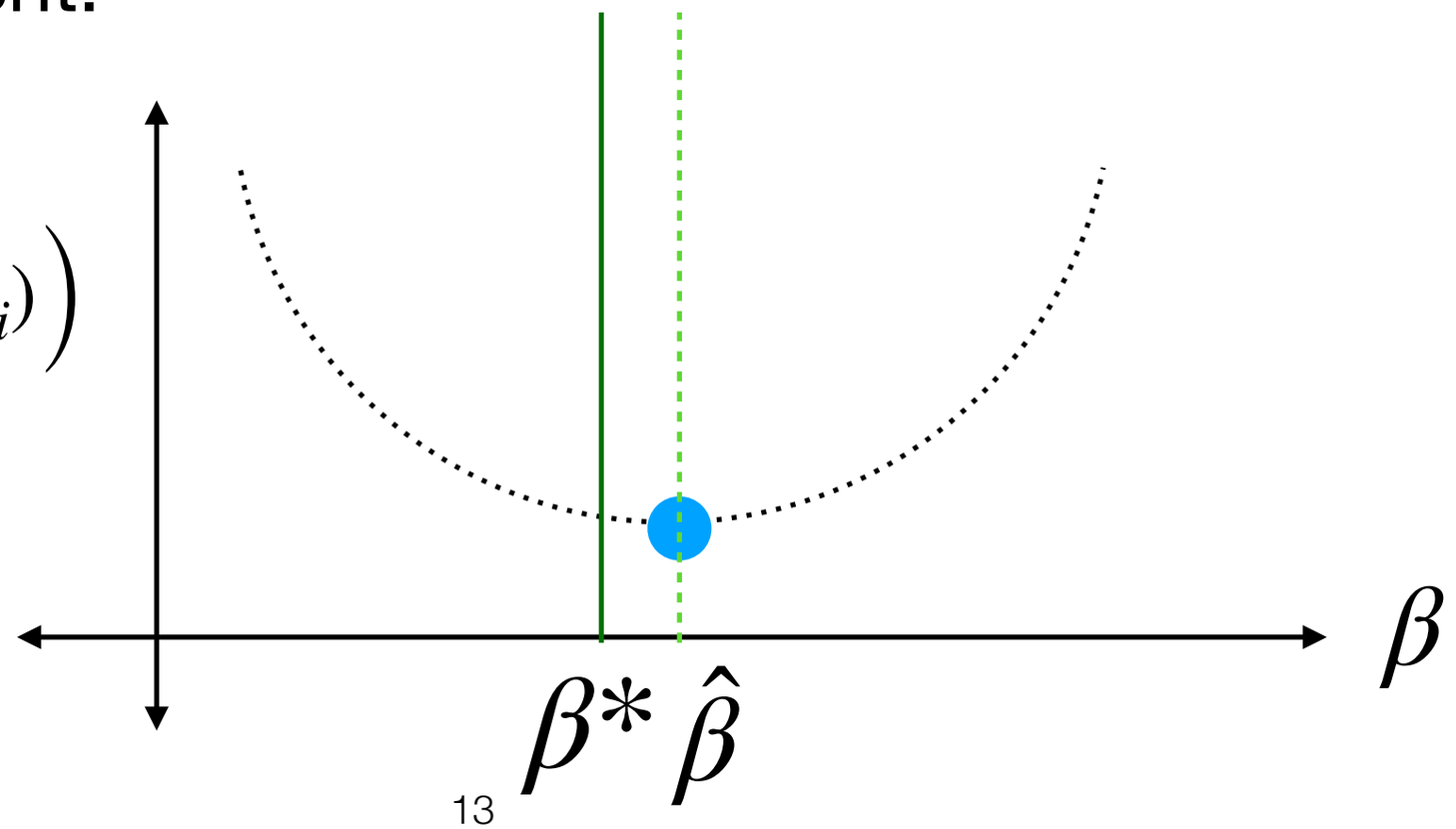
Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent:

$$\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)$$



Solving the optimization problem

- How do you find a β that solves the optimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, f_\beta(x_i) \right)?$$

- You could use a general optimization algorithm like gradient descent.

Many machine learning algorithms can be written as the minimizer of the average loss and solved using general-purpose optimization algorithms.

The special case of linear regression

- Okay, for linear regression, there is actually a closed-form solution. We won't go over the equations here.

- Using a closed-form solution means you can fit a linear model very quickly, much faster than using a general optimization procedure.

Today's lecture

1. Linear regression as an optimization procedure
2. **Classification**
    1. **Loss functions**
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. K-nearest neighbors
    6. Receiver Operating Characteristic (ROC)

Classification

- Supervised learning where the outcome is a class variable
- Unordered class variable such as:
  - `dx` $\in$ {`cancer`, `benign`}
  - `blood_type` $\in$ {`typeA`, `typeB`, `typeAB`, `typeO`}
- Given a set of predictors X and a set of possible classification outcomes $\mathscr{C}$, the objective is to find a function/algorithm $f(X)$ that takes X as input and predicts the outcome Y or the probability of each outcome.

Classification

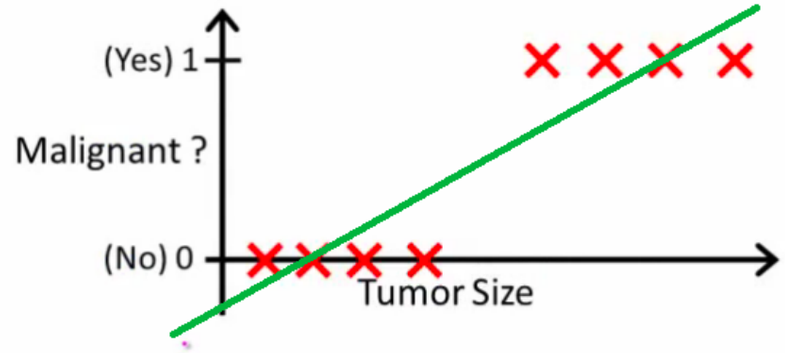
- **Q**: Can we treat this as a regression problem?

- **A**: You probably don't want to do that for a number of reasons:

  - Regression models can predict values outside the range of possible values.

  - If you have multiple classes and you want to make this a regression problem, you'll need to order the classes. This might not make sense for things that are not naturally ordered.

  - Even if things are ordered, it can be hard to assign numerical values, because that will imply how far apart the outcomes are. For example, should we actually code cancer stage using the original numbers?



Type O … 0
Type A … 1
Type B … 2
Type AB … 3? 4?

Classification loss functions

- **0-1 loss**: Is the predicted outcome different from the observed outcome?
  - $\ell(y, \hat{y}) = 1\{\hat{y} \neq y\}$

While intuitive, this is a really difficult optimization problem. We tend to use this for testing, not for training

- **Negative log likelihood (cross-entropy)**: Is the observed outcome unlikely under the given probability model?
  - $\ell(y, \hat{p}) = -y \log \hat{p} - (1 - y)\log(1 - \hat{p})$

Mathematically, much easier to work with. Commonly used to train models.

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. **Logistic Regression**
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. k-Nearest Neighbors
    6. Receiver Operating Characteristic (ROC)

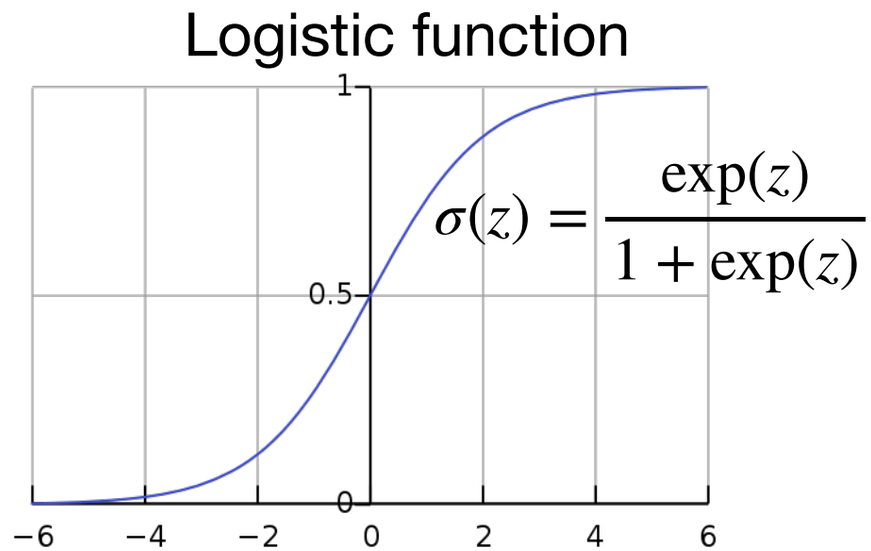
Logistic regression

- For binary classification settings where $Y \in \{0,1\}$

- Logistic regression assumes that the probability that $Y = 1$ for variable X has the form

$$\Pr(Y = 1 \mid X) = \underbrace{\frac{\exp(X^\top \beta + \beta_0)}{1 + \exp(X^\top \beta + \beta_0)}}.$$

This value is always between 0 and 1 for any value of β_0 and β



- Recall that the coefficients β can be interpreted as the log odds ratio.

Logistic regression as an optimization problem

To perform logistic regression, we do **maximum likelihood estimation,** which is equivalent to minimizing the negative log likelihood loss:

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f_{\beta_0,\beta}(x_i)\right)$$

Negative log likelihood: How unlikely is it to observe this data given this probability model?

Predicted probability

Find a probabilistic model such that the observed data is highly probable

where

$$\ell(y, f(x)) = -y \log f(x) - (1-y)\log(1-f(x))$$

and $f_{\beta_0,\beta}(x)$ is the probability of $Y = 1$ for input x:

$$f_{\beta_0,\beta}(x) = \frac{\exp(x^\top \beta + \beta_0)}{1 + \exp(x^\top \beta + \beta_0)}.$$

Example: Predicting cause of death

- UCSF's POST-SCD study has autopsied every incident out-of-hospital sudden death within San Francisco.

- Can we identify biomarkers in the heart that predict whether a patient died of a sudden arrhythmic death versus trauma?

- $Y \in \{\text{trauma, sudden arrhythmic death}\}$

$$Y = 0 \qquad\qquad Y = 1$$

- $X =$ standardized heart weight

Is heart weight associated with cause of death?

```
logistic_m <- glm(y ~ x, family = "binomial")
summary(logistic_m)
```

```
Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9761  -0.9510  -0.5801   1.0430   1.9822

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2626     0.2222  -1.182  0.23728
x             0.9688     0.2576   3.760  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confidence intervals

- We can construct confidence intervals to quantify our uncertainty about the log odds ratio:

```
> confint(logistic_m)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) -0.7043626 0.1715319
x            0.4993661 1.5173068
```

- Interpretation: “A one unit increase in the standardized heart weight is associated with an increase in the log odds by 0.97 (CI: 0.50, 1.51).”

Getting predictions from the model

- What are the probability that the cause of death was sudden arrhythmic death for standardized heart weight of 1?

```
> predict(logistic_m, newdata = data.frame(x=c(1)), type = "response")
        1
0.6682795
```

Wow, what a small p-value!

Q: Does a small p-value mean that my logistic regression model is a really good prediction model?

- **A**: Not necessarily
- **B**: Yes! Time to celebrate!

```
Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9761  -0.9510  -0.5801   1.0430   1.9822

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2626     0.2222  -1.182  0.23728
x             0.9688     0.2576   3.760  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A p-value indicates how likely it is to observe this data if the outcome is not associated with the variable. If the association is very weak, you can still get a large p-value by increasing the size of the dataset. However, that doesn’t mean you have a very accurate prediction model!

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
3. **Multinomial Regression**
4. Linear Discriminant Analysis
5. k-Nearest Neighbors
6. Receiver Operating Characteristic (ROC)

Multinomial regression

- Sometimes the outcome can belong to one of K classes. Instead of doing logistic regression, we can do multinomial regression.

- The model is indexed by parameters $(\beta_{0,k}, \beta_k)$ for $k = 1,...,K$ the form. The probability of the kth class for input x is defined as

$$\Pr(Y = k \mid X = x) = \frac{\exp(x^\top \beta_k + \beta_{0,k})}{\sum_{l=1}^{K} \exp(x^\top \beta_l + \beta_{0,l})}.$$

(also known as the soft-max function)

(value between 0 and 1)

- To fit a multinomial regression model, we also perform maximum likelihood estimation.

Little quiz!

- **Q**: How do logistic and multinomial regression try to model the data?

- **A**: They try to model Y as a function of X

- **B**: They try to model X as a function of Y

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
3. Multinomial Regression
4. **Linear Discriminant Analysis**
5. k-Nearest Neighbors
6. Receiver Operating Characteristics (ROC)

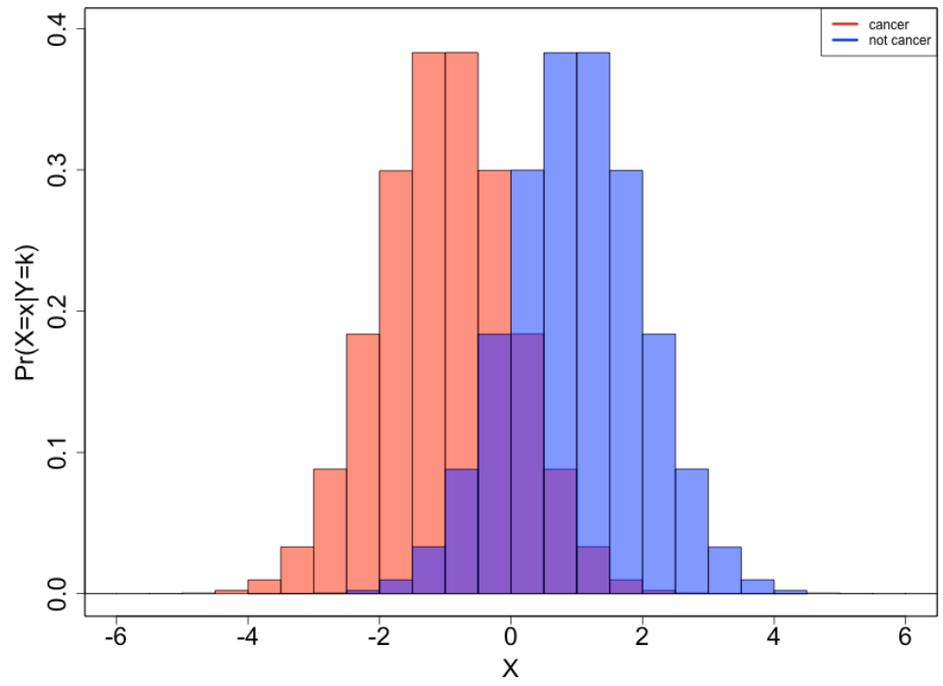
Modeling X|Y instead of Y|X

- Up to now, we've been modeling $\Pr(Y = y \mid X = x)$.
- How about switching things around and modeling $\Pr(X = x \mid Y = y)$?

Modeling X|Y instead of Y|X

When do we want to model X|Y vs Y|X? Possible reasons:

1. You don't have a lot of data, but you have a good understanding of X|Y. For example, you know that X is normally distributed within each category.



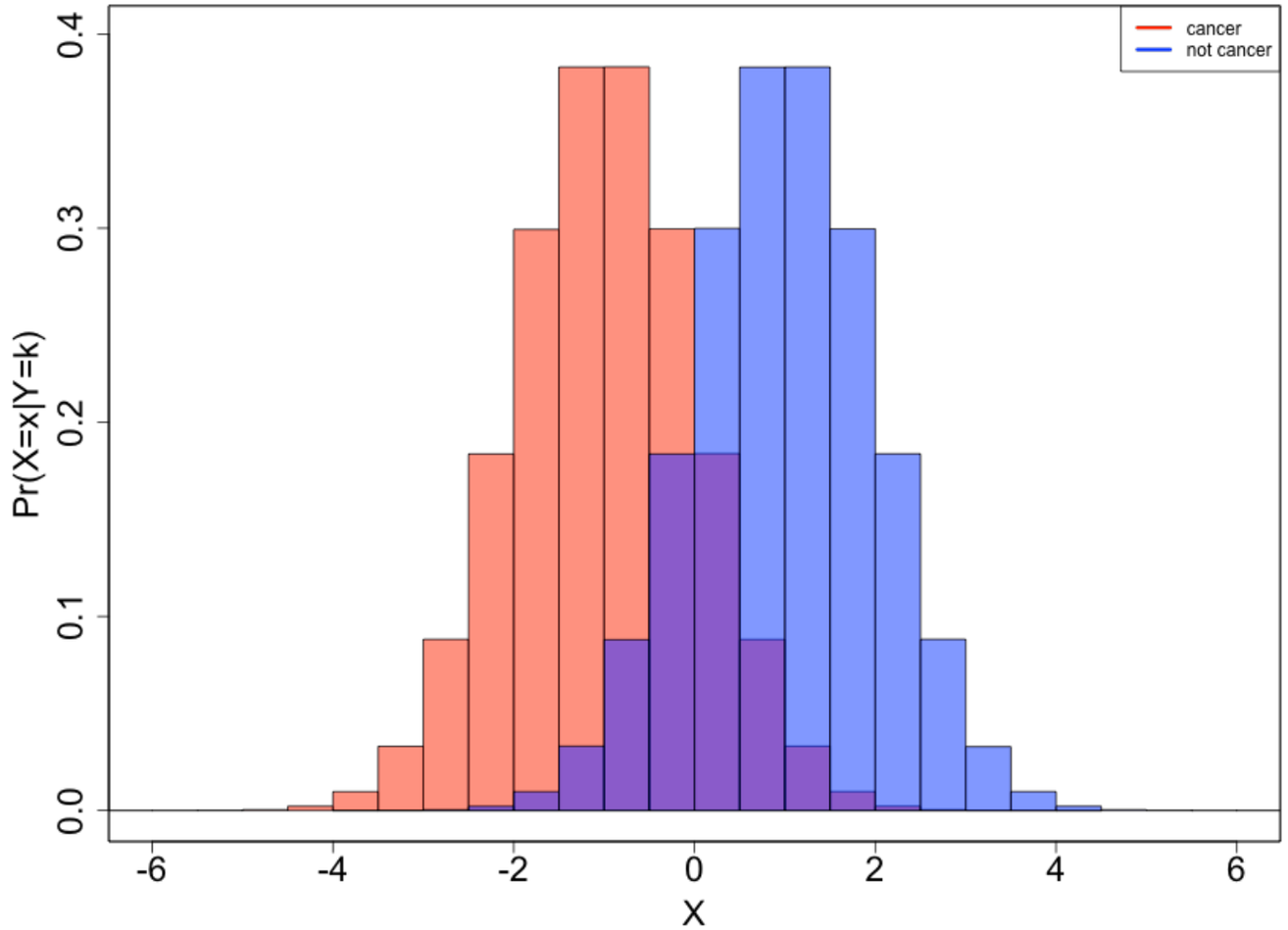
2. You want to generate examples of X|Y. For example, faces of men versus women.

Linear discriminant analysis (LDA)

- The primary motivation for using LDA is that you already have a model in mind for $\Pr(X = x \mid Y = y)$.

(Highly) Parametric model!

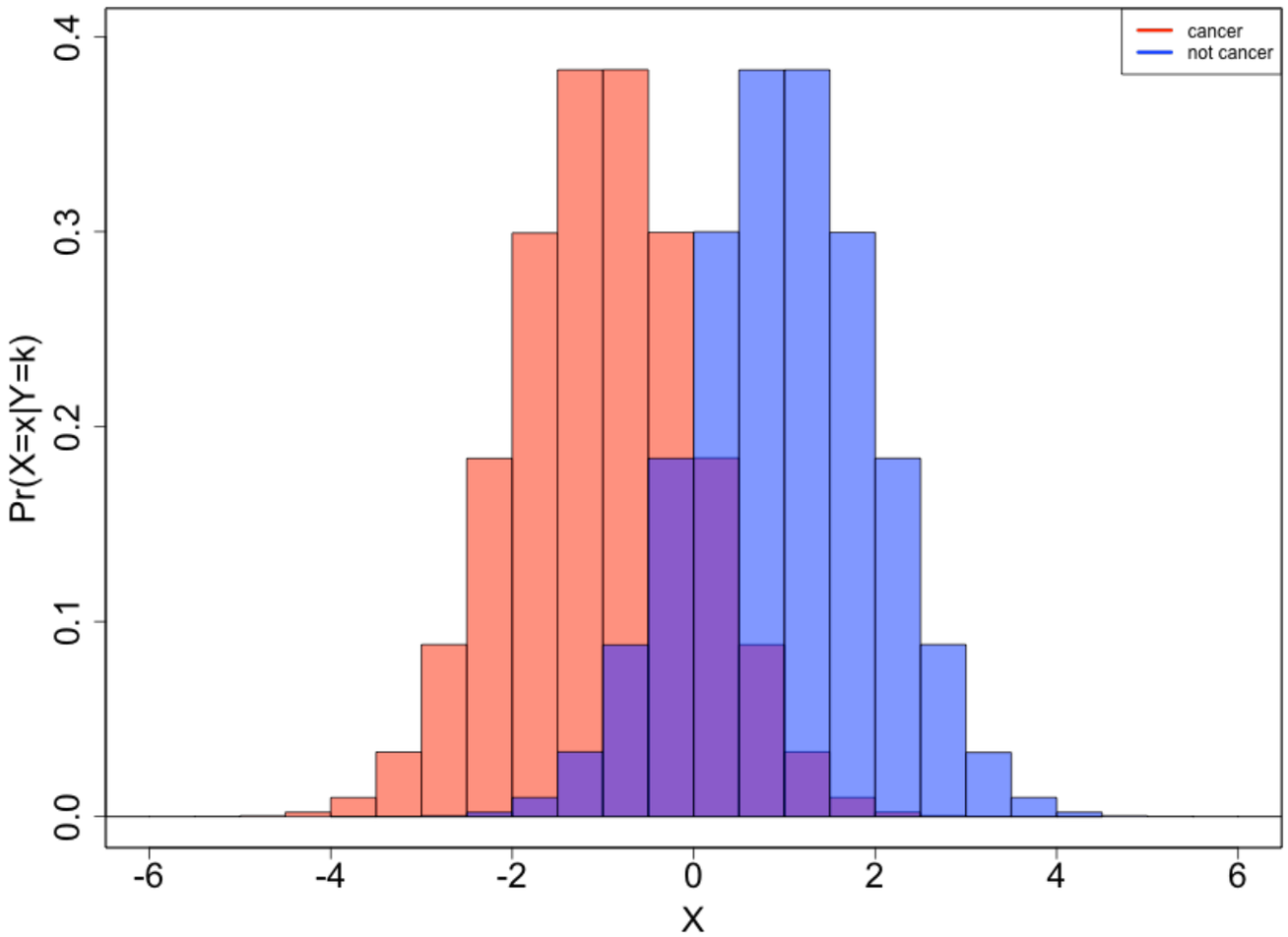
- In LDA, $X \mid Y$ is assumed to be a normal distribution:



Linear discriminant analysis

To model $\Pr(X = x \mid Y = y)$,
LDA estimates:

- The mean of X within each class
- The variance of X, which is usually assumed to be the same across all classes.



- Also, we model the probability of each class $\Pr(Y = y)$

Bayes' Theorem

- How do you predict Y given X if you only have a model of X given Y?

- Bayes' theorem states that

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$

$$= \frac{\Pr(X = x \mid Y = y)\Pr(Y = y)}{\sum_{y'=0}^{1} \Pr(X = x \mid Y = y')\Pr(Y = y')}$$

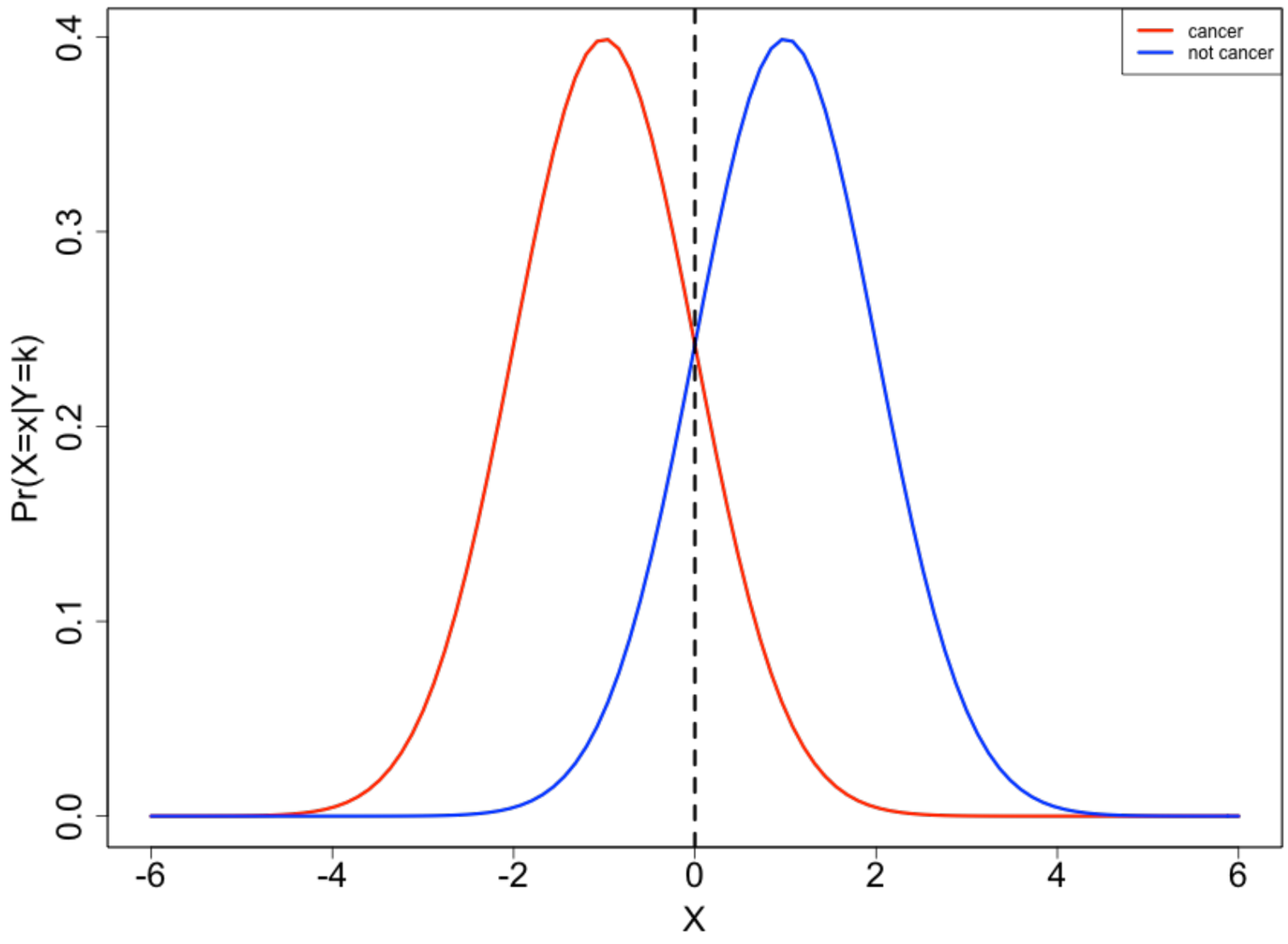
Model for the prevalence of Y

Model of X|Y

Linear discriminant analysis

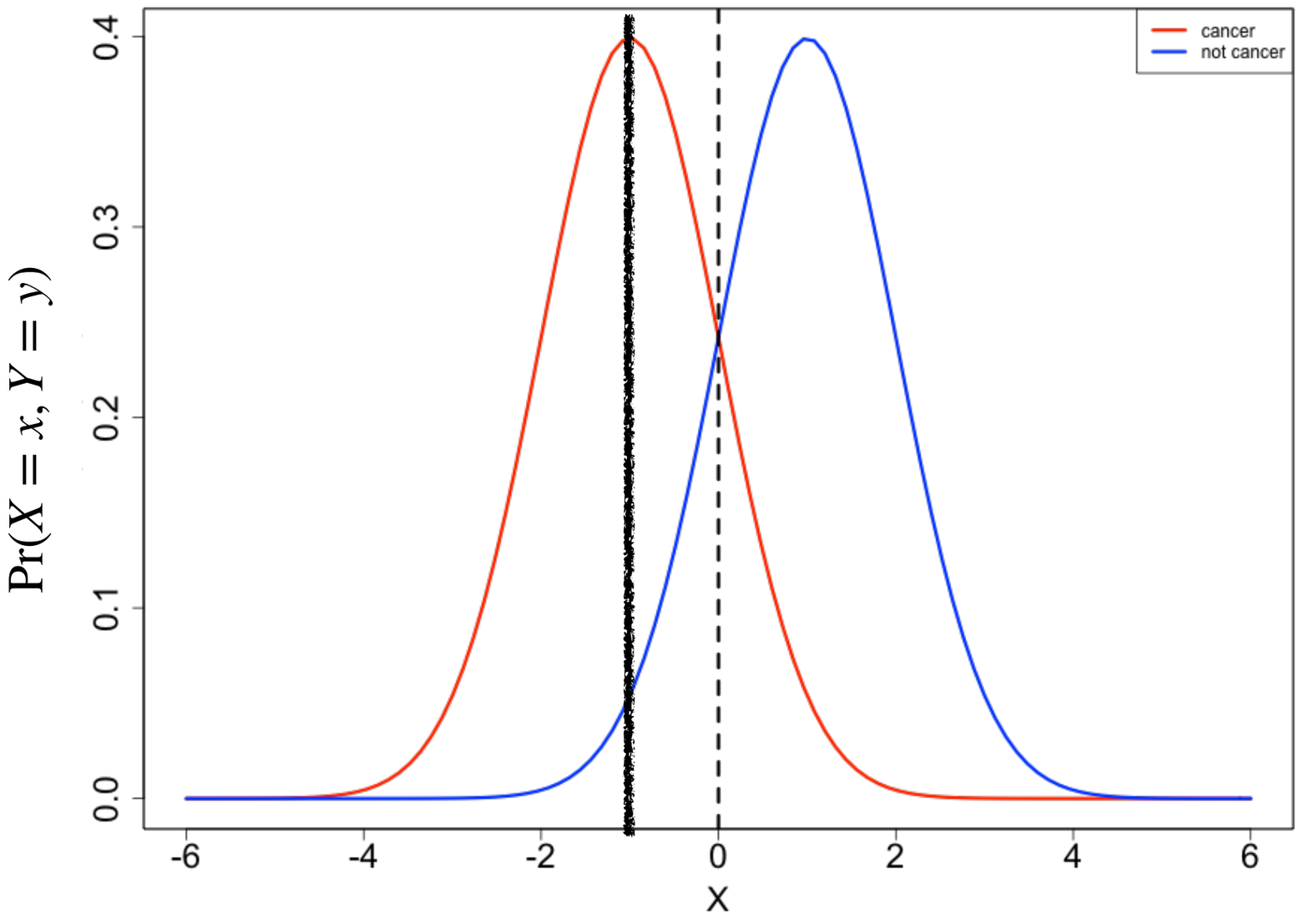
- Given the fitted model for $\Pr(X = x \mid Y = y)$, which class is more likely for a subject with $X = -1$?

To answer this, you need to know the prevalence of the two diseases.



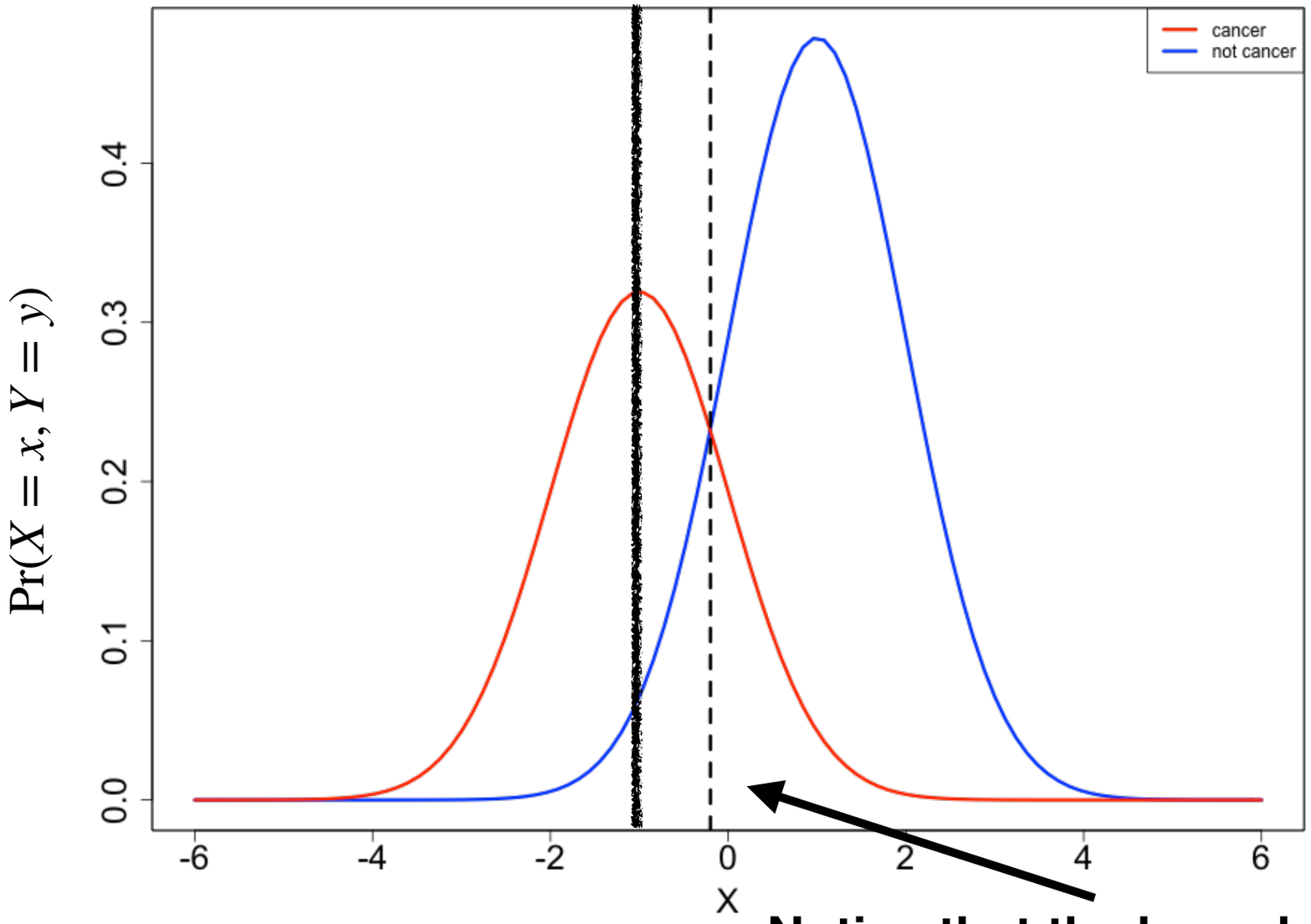
Linear discriminant analysis

- Suppose we know that $\Pr(Y = 1) = \Pr(Y = 0)$. After calculating $\Pr(X = x, Y = y)$, do we know which class is more likely for a subject with $X = -1$?



Linear discriminant analysis

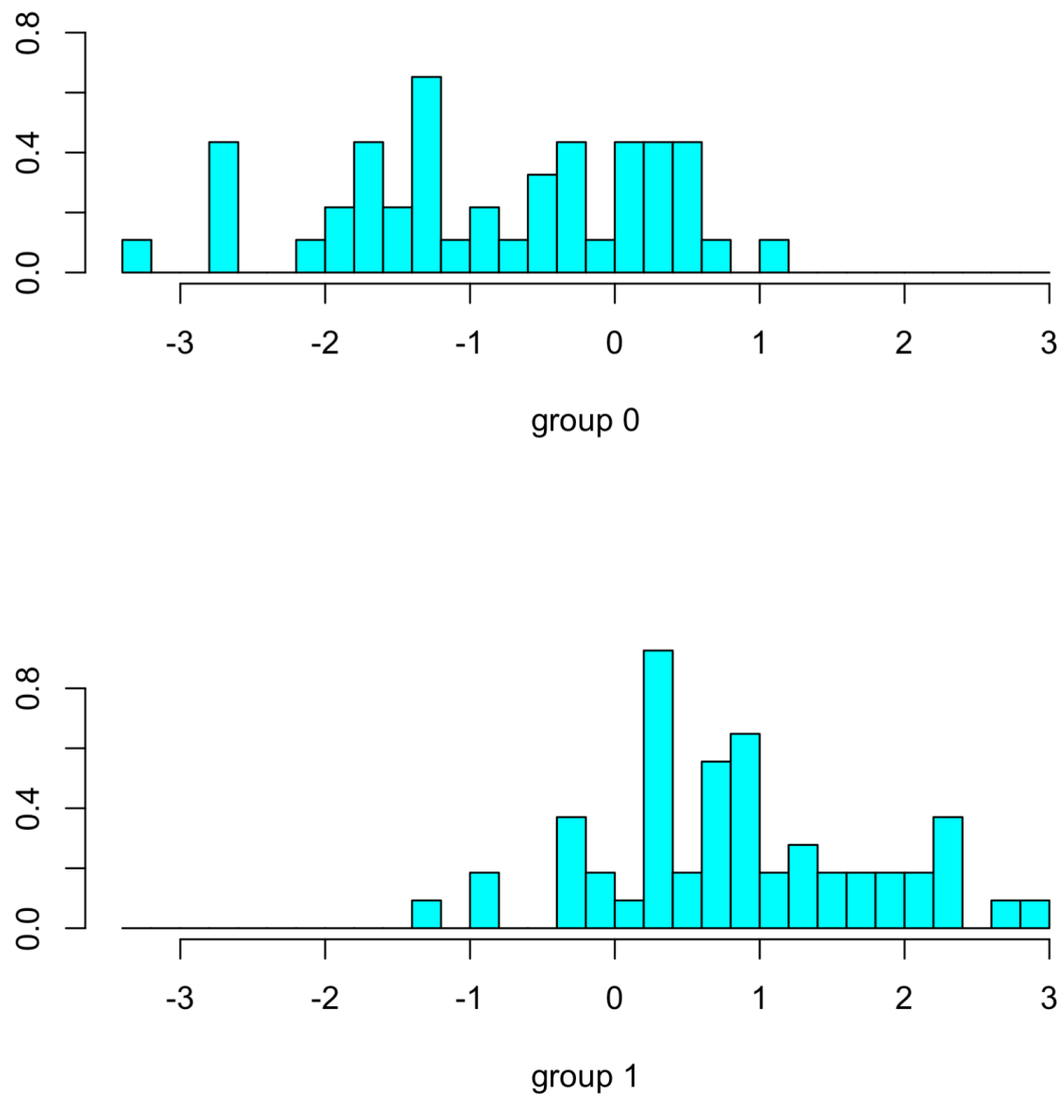
- Suppose instead that $\Pr(Y = 0) = 0.2$ and $\Pr(Y = 1) = 0.8$. We recalculate $\Pr(X = x, Y = y)$. Which class is more likely for a subject with $X = -2$?



Notice that the boundary shifted a bit!

Post-SCD example, continued.

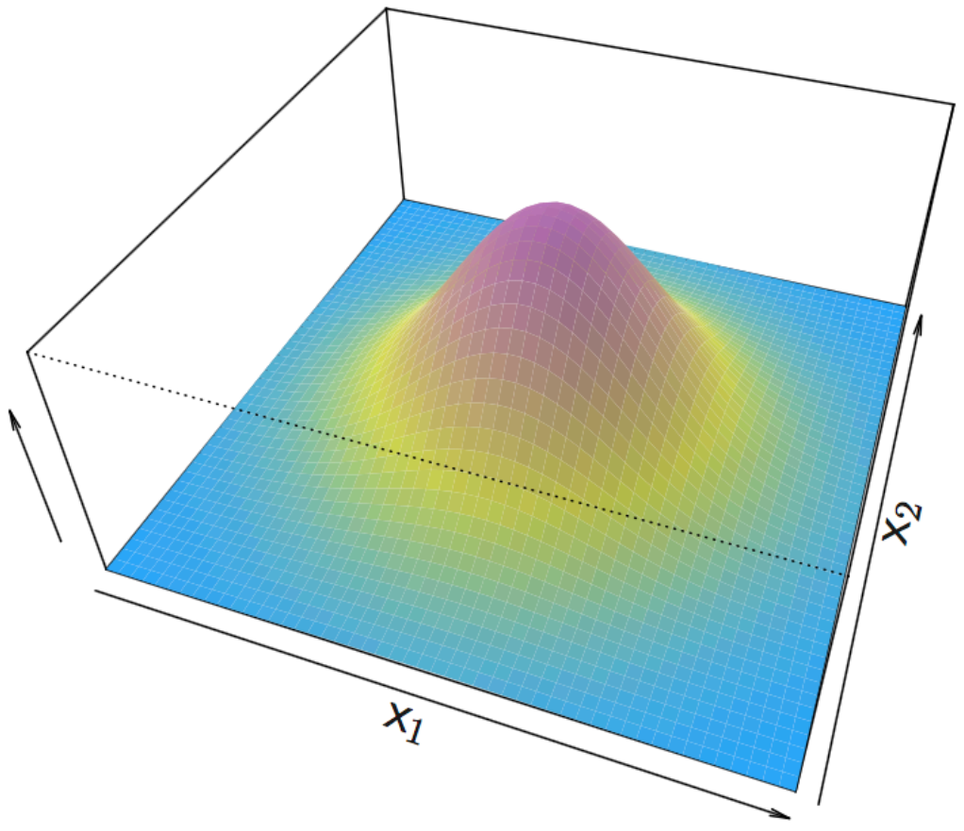
```
> lda_m <- lda(y ~ x)
> lda_m
Call:
lda(y ~ x)

Prior probabilities of groups:
    0     1
0.46 0.54

Group means:
            x
0 -0.7883936
1  0.5240000

Coefficients of linear discriminants:
       LD1
x 1.279236
```



Multiple variables in LDA

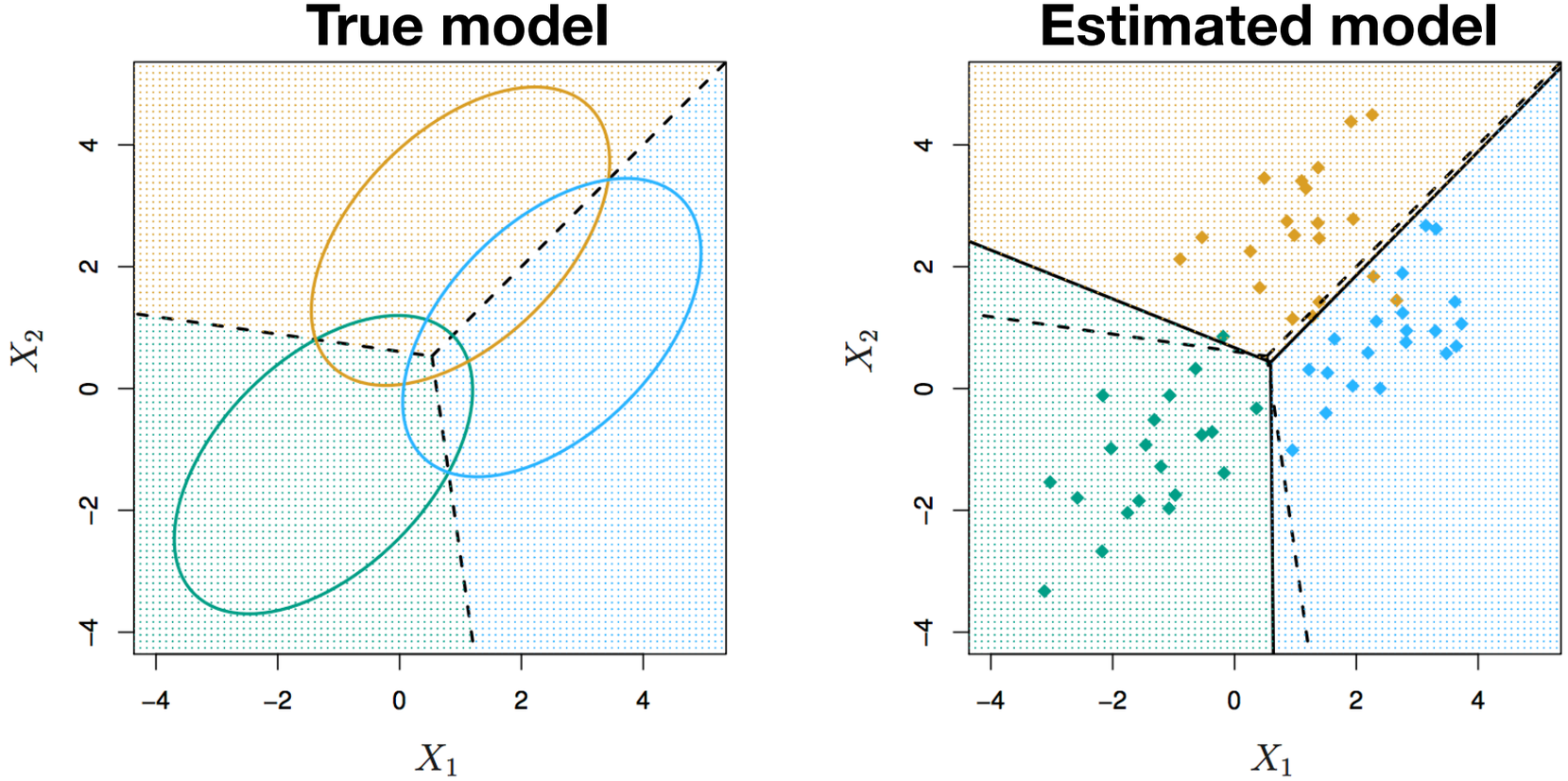
- When there are multiple predictors, LDA fits a Gaussian distribution for each class.

- Note that it again assumes that the shape of the Gaussian distributions are the same across all classes.



Decision boundaries in LDA

- If we are forced to predict a single class for each point, we should draw ***linear*** boundaries between the Gaussian distributions learned by LDA to minimize the 0-1 loss (aka the Bayes optimal classifier). Hence the name.



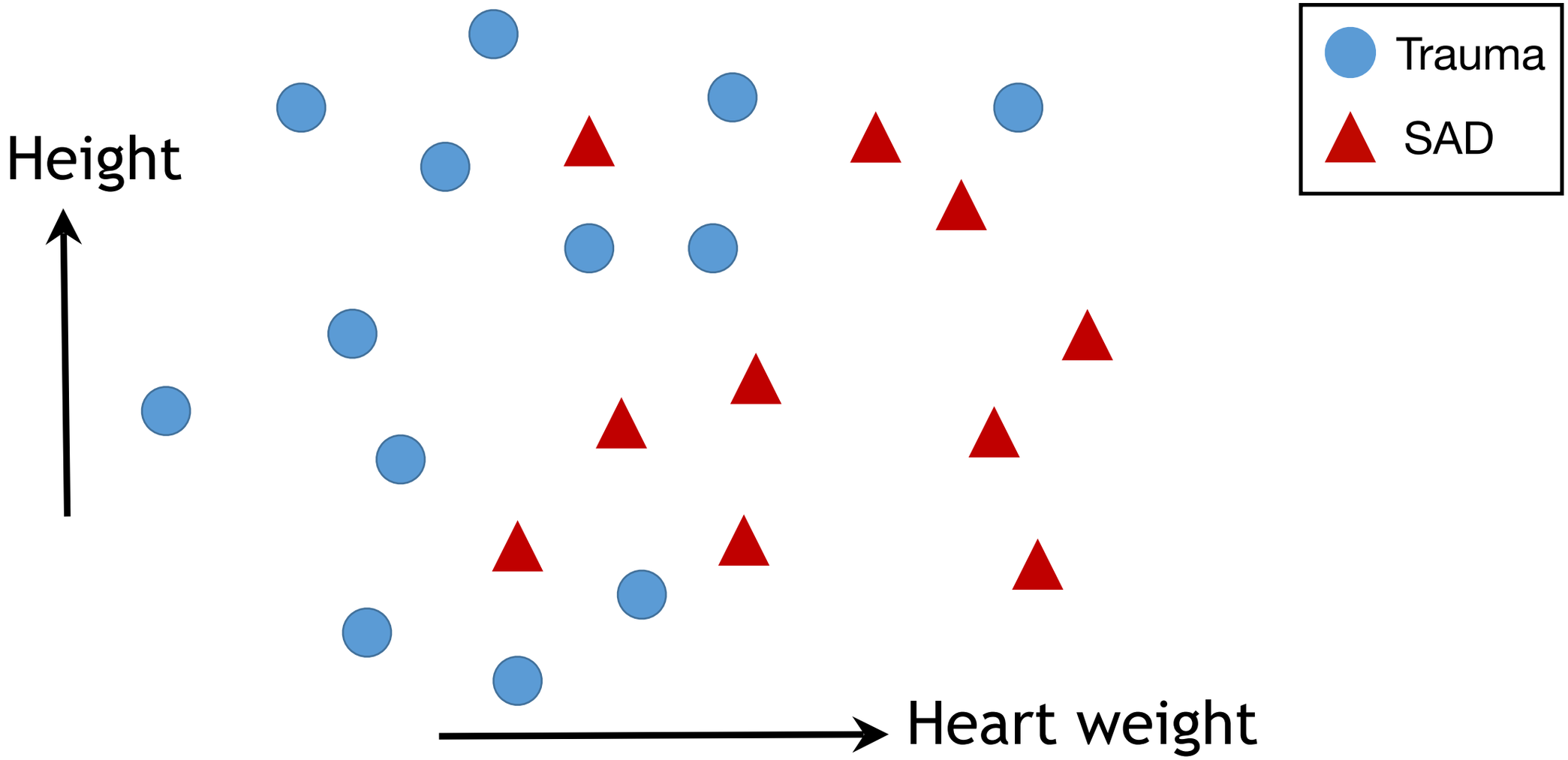
Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis

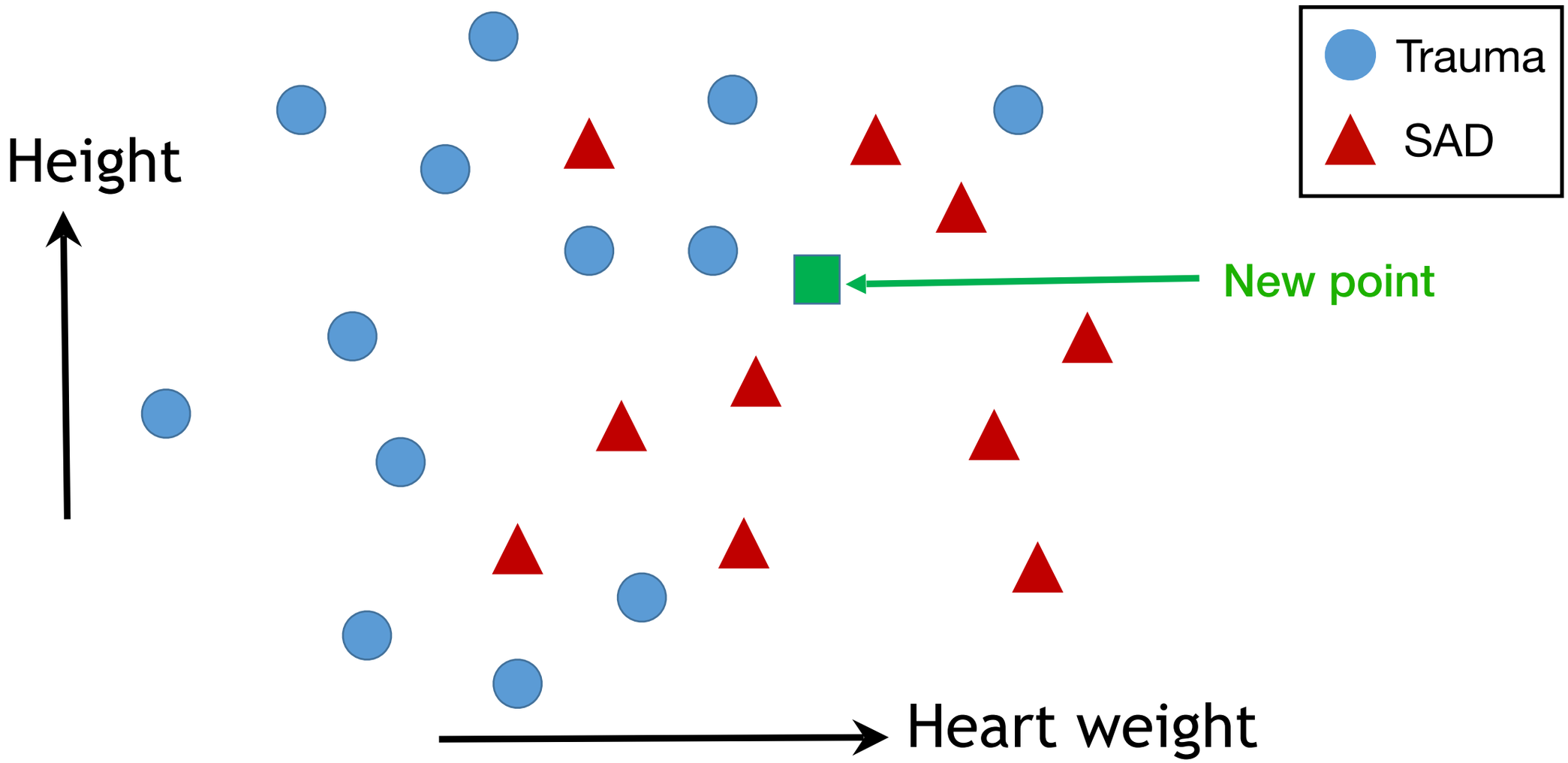
    } Parametric

    5. **k-Nearest Neighbors** Non-parametric
    6. Receiver Operating Characteristic (ROC)

K-nearest neighbors

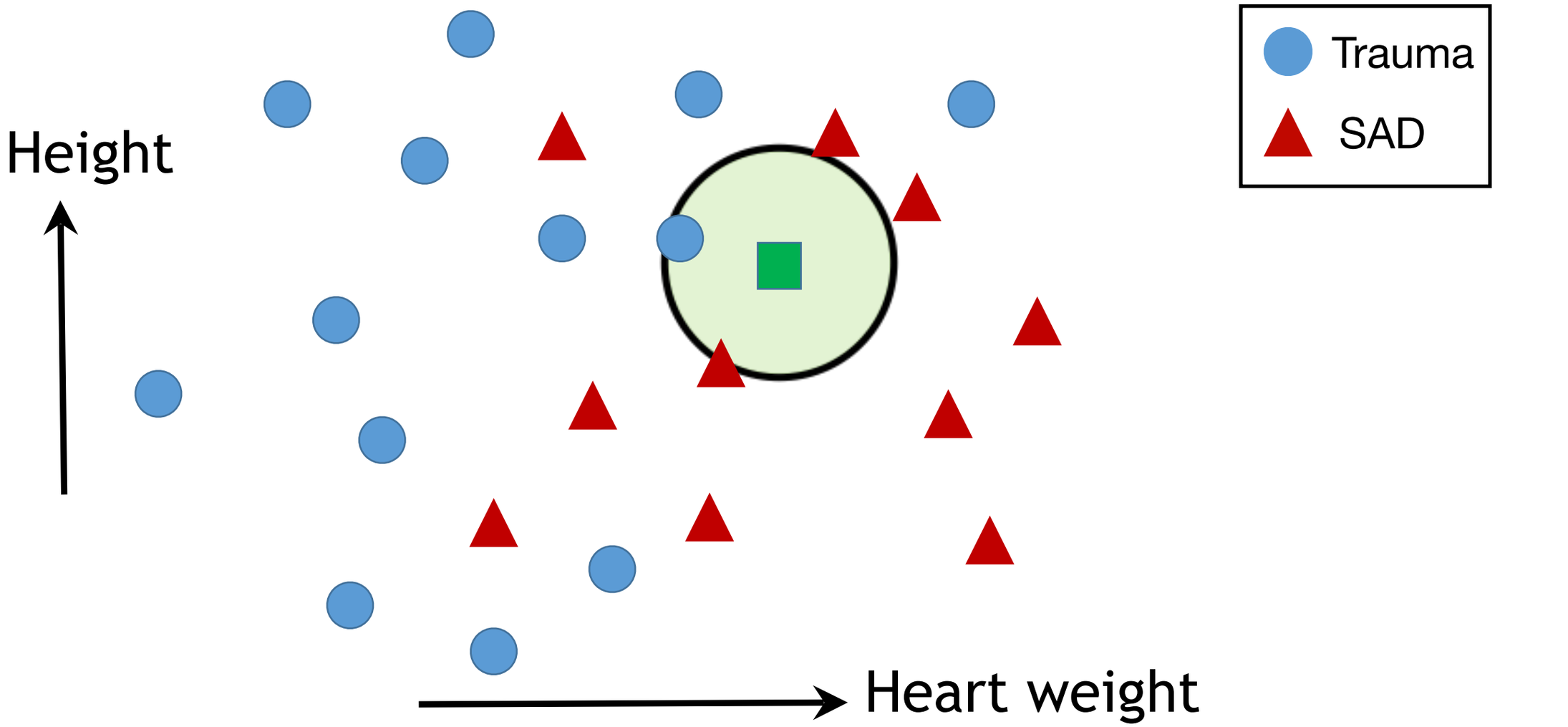


K-nearest neighbors



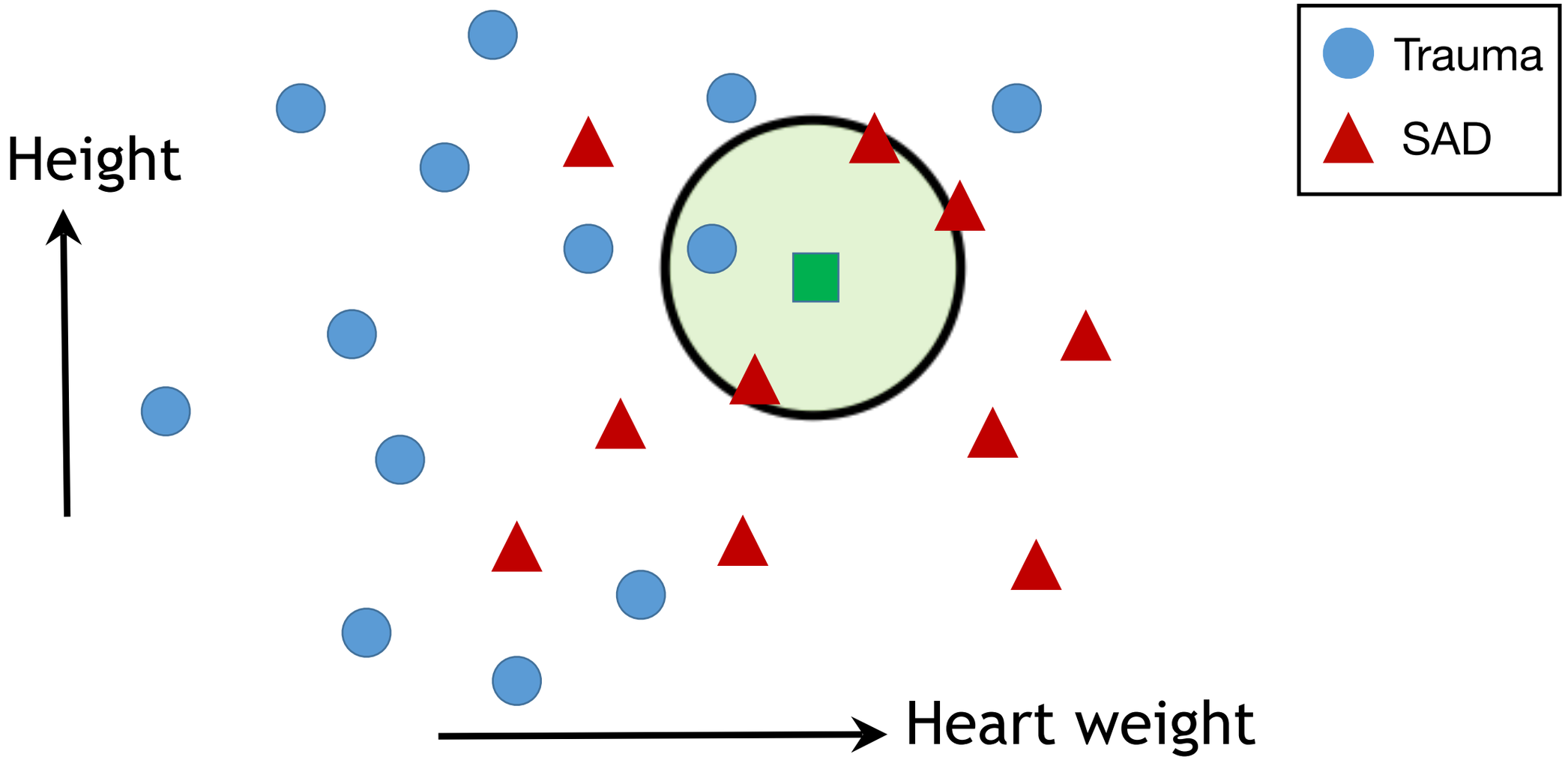
K-nearest neighbors, K=1

Closest neighbor is Trauma.



K-nearest neighbors, K=3

2/3 closest neighbors are SAD.

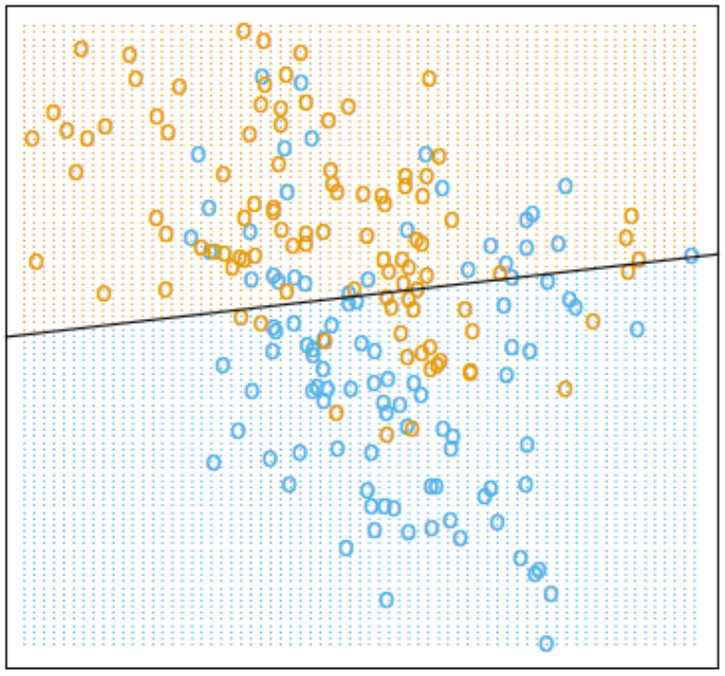


K-Nearest Neighbors

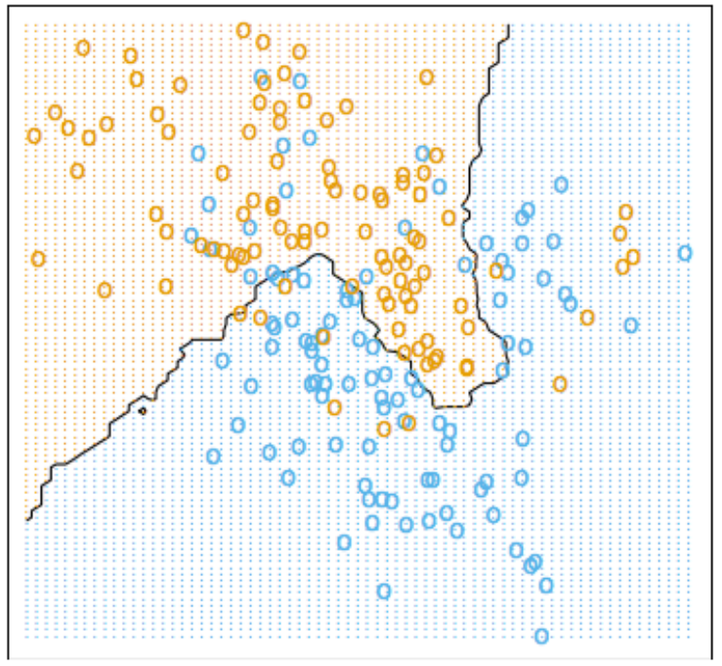
- Take majority vote among k-nearest neighbors. Ties are broken at random.
- Works for problems with more than 2 classes
- The method is very sensitive to how the distance metric is defined.
- If all the predictors are continuous, common choices are the Euclidean distance and Pearson/Spearman correlation. (You probably want to standardize predictors as well.)
- If you have categorical predictors, you'll need to decide how distance is defined between the categories.

		Blood type			
		A	B	AB	O
Blood type	A	0	2	1	3
	B		0	1	3
	AB			0	4
	O				0

Decision boundaries from LDA vs K-nearest neighbors



LDA

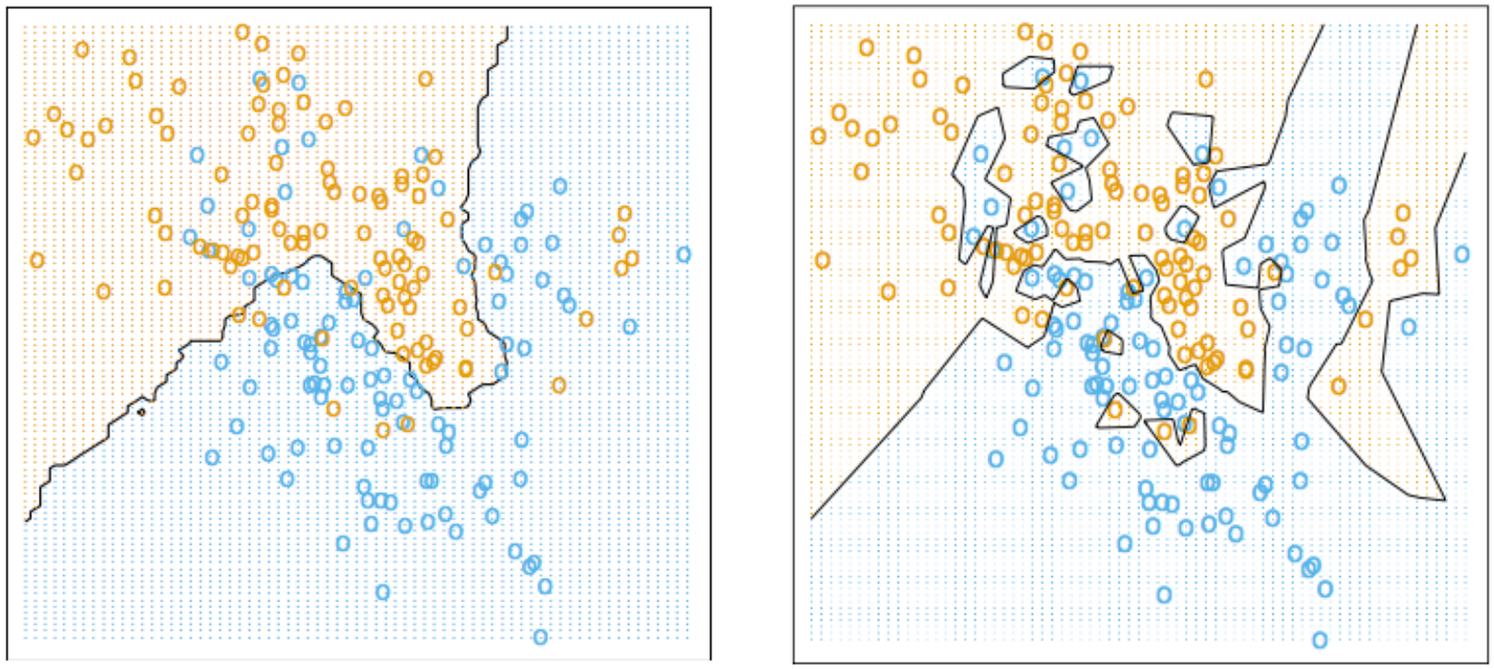


15-NN

When choosing between models, we must take into account the bias-variance tradeoff!

Bias-variance tradeoff

Q: How do we pick K?



15-NN **1-NN**

A: Training-test split. Evaluate the test error for different values of K, and pick the K with the smallest test error.

We'll talk about more sophisticated ways for selecting K next lecture.

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. k-Nearest Neighbors
    6. **Receiver Operating Characteristic (ROC)**

Ways to evaluate a classification model…

- We've talked about 0-1 loss and the negative log likelihood.

- In addition, people often use sensitivity and specificity.

Sensitivity and Specificity

Truth

Prediction	**Negative**	**Positive**	Total
Negative	True Negative TN	False Negative FN	FN+TN
Positive	False Positive FP	True Positive TP	TP+FP
Total	FP+TN	TP+FN	N

Specificity = $\frac{TN}{TN + FP}$ =
Proportion of true negatives correctly identified as negative

Sensitivity = $\frac{TP}{TP + FN}$ =
Proportion of true positives correctly identified as positive

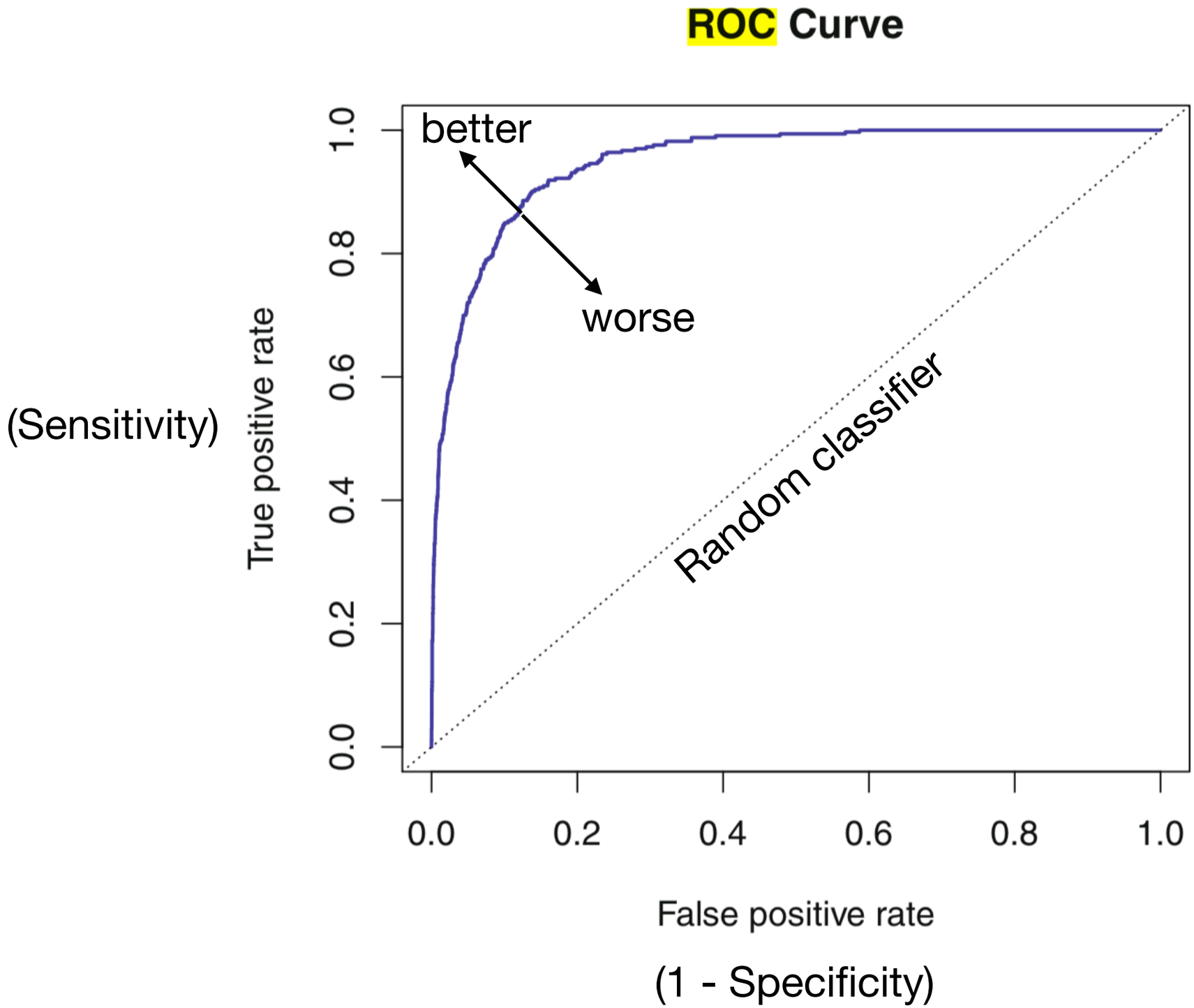
Thresholding predictions

- In practice, many classification models output a continuous value, such as a probability.

- In this case, you need to pick a threshold for deciding positive versus negative tests. Only then can you calculate sensitivity and specificity.

- Rather than calculating sensitivity and specificity for a single threshold, we could plot the two values *for all possible thresholds*.

Receiver operating characteristic (ROC)

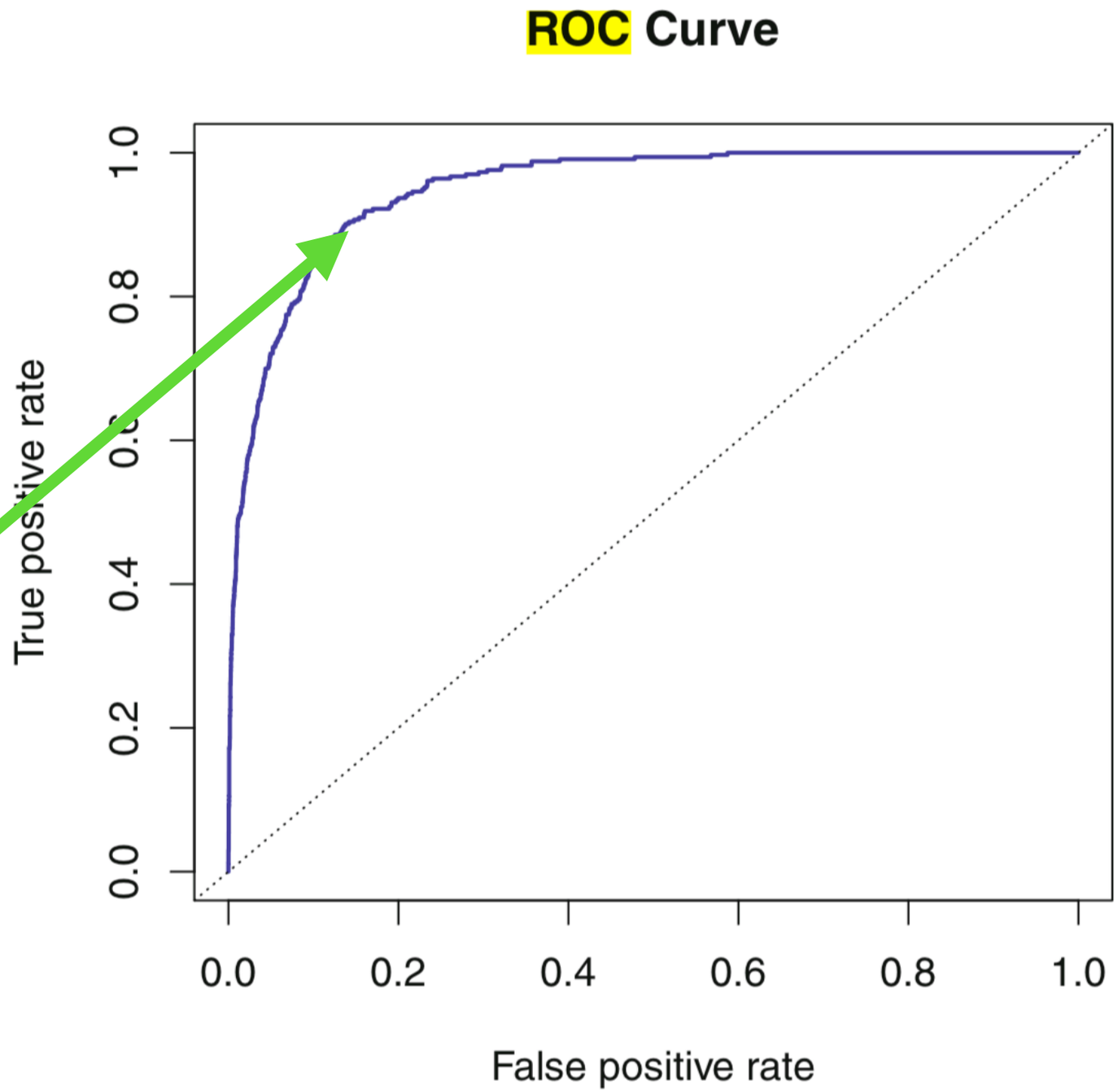


Picking a threshold

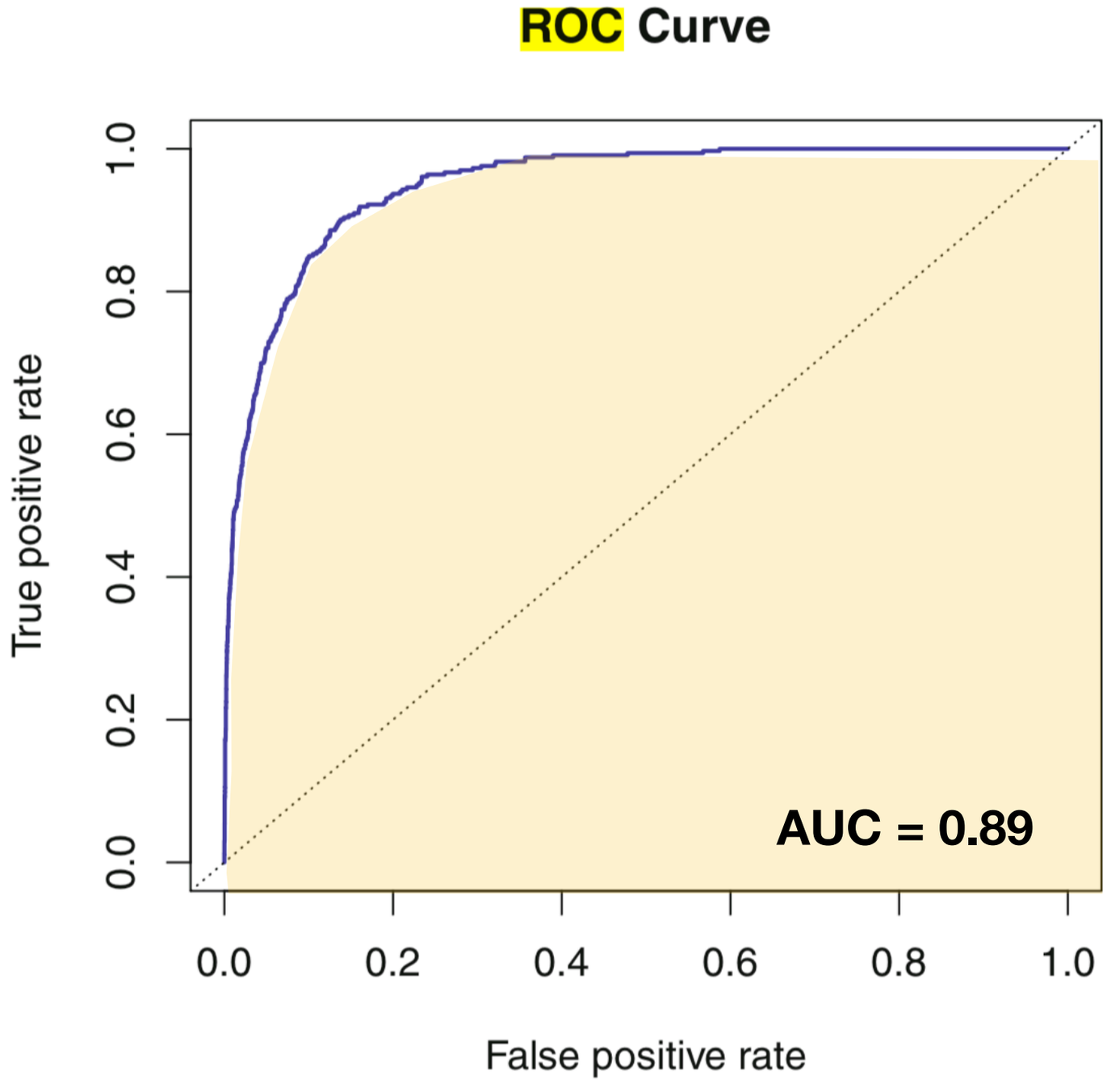
The ROC can help you pick a threshold for a binary classifier. For example:

- Pick the threshold that achieves the largest TPR given a maximum FPR of 0.2

- Pick the elbow (the farthest point from the diagonal line). This treats false positives and false negatives equally.



Area under the Receiver operating characteristic (AUROC or AUC)



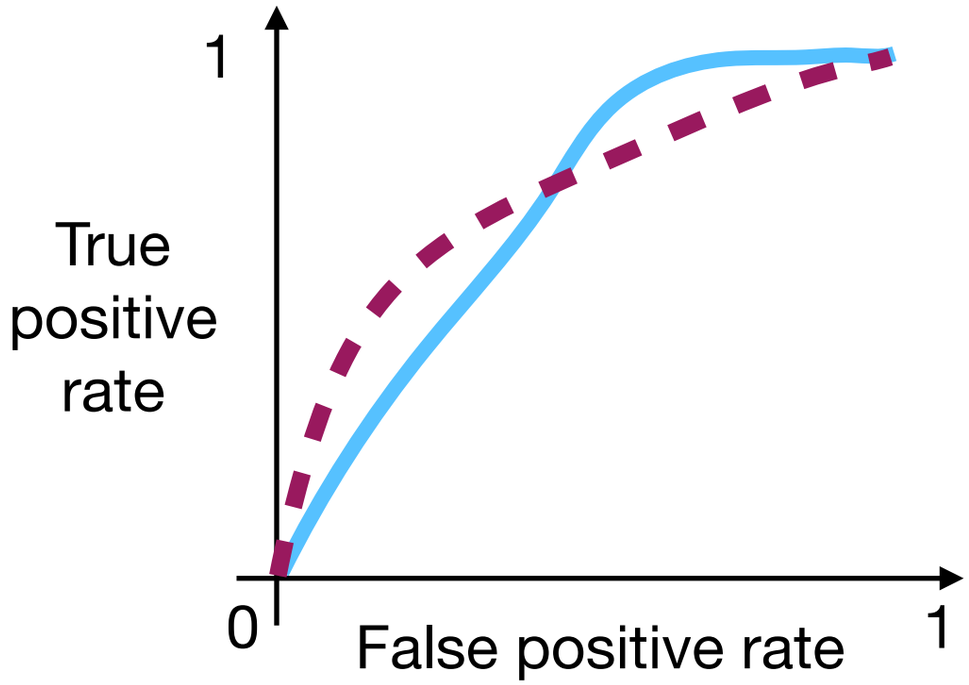
- AUC of a random classifier is 0.5. Ideal AUC is 1.
- AUC characterizes ***model discrimination***, i.e. how well the model separates positive versus negative cases.
- Describes the behavior of a model independent of disease prevalence.

Comparing two classification models

- Suppose we are given two models that predict a patient's risk of sudden arrhythmic death. We must pick one of the two models. To avoid issues of alarm fatigue, we plan to use deploy the prediction model so that its false positive rate is 20%. Which of the two models would you pick?

- **A:** Model with the purple, dashed ROC curve

- **B:** Model with the blue, solid ROC curve



Assessing the Performance of Prediction Models

A Framework for Traditional and Novel Measures

Ewout W. Steyerberg,[a] *Andrew J. Vickers,*[b] *Nancy R. Cook,*[c] *Thomas Gerds,*[d] *Mithat Gonen,*[b] *Nancy Obuchowski,*[e] *Michael J. Pencina,*[f] *and Michael W. Kattan*[e]

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	*c* statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB)	Cross-table	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)
	Decision curve analysis (DCA)	Decision curve	

Today's lecture

1. Linear regression as an optimization procedure
2. Classification
    1. Loss functions
    2. Logistic Regression
    3. Multinomial Regression
    4. Linear Discriminant Analysis
    5. k-Nearest Neighbors
    6. Receiver Operating Characteristic (ROC)