BIOST 546: Machine Learning for Biomedical Big Data

Ali Shojaie

Lecture 4: Classification for Biomedical Big Data - Part I
Spring 2017

Recap

- Bias-variance tradeoff & Training/Test error
- Cross validation and related procedures
- Methods for reducing model complexity:
  - subset selection (variable pre-selection, best subset selection, forward step-wise selection)
  - regularization (ridge and lasso)
  - dimension reduction (PCR and PLS)

Today & Next Lecture

- High dimensional classification:
  - Classification using linear regression
  - Logistic regression (& penalized logistic regression)
  - KNN classification
  - LDA & QDA
  - Support Vector Machines (SVM)

Class Projects

- 40% of the total grade

Class Projects

- 40% of the total grade
- Ideally on applications of statistical machine learning to biomedical big data
  - Typical project includes reproducing the results of a paper that uses statistical learning to analyze biomedical big data, or applying new methods to previously analyzed, publicly available dataset.
  - You may need to preprocess the data and do preliminary analyses to remove batch effects etc (see BIOST 544, 545).
  - Please post your proposal as a discussion by the *end of next week*. If you don't have a topic in mind, I will try to match you with others.
  - There will be groups of 2 (at most 3) working on the same project, you can team up yourselves, but I may suggest changes.

Class Projects

- 40% of the total grade
- Ideally on applications of statistical machine learning to biomedical big data
  - Typical project includes reproducing the results of a paper that uses statistical learning to analyze biomedical big data, or applying new methods to previously analyzed, publicly available dataset.
  - You may need to preprocess the data and do preliminary analyses to remove batch effects etc (see BIOST 544, 545).
  - Please post your proposal as a discussion by the *end of next week*. If you don't have a topic in mind, I will try to match you with others.
  - There will be groups of 2 (at most 3) working on the same project, you can team up yourselves, but I may suggest changes.
- Tentative schedule:
  - Finalize the project topic: Thursday April 27th
  - Proposal presentations: 5%
  - 2-page project proposal & progress report: 5%
  - Final Presentations: 10%
  - Full report (max 7 pages in ICML format (TBD)): 20%

Classification

- Regression involves predicting a continuous-valued response, like tumor size.

Classification

- Regression involves predicting a continuous-valued response, like tumor size.
- Classification involves predicting a categorical response:
  - Cancer versus Normal
  - Tumor Type 1 versus Tumor Type 2 versus Tumor Type 3

Classification

- Regression involves predicting a continuous-valued response, like tumor size.
- Classification involves predicting a categorical response:
  - Cancer versus Normal
  - Tumor Type 1 versus Tumor Type 2 versus Tumor Type 3
- Classification problems tend to occur even more frequently than regression problems in the analysis of biomedical data.

Classification

- Regression involves predicting a continuous-valued response, like tumor size.
- Classification involves predicting a categorical response:
  - Cancer versus Normal
  - Tumor Type 1 versus Tumor Type 2 versus Tumor Type 3
- Classification problems tend to occur even more frequently than regression problems in the analysis of biomedical data.
- Just like regression,
  - Classification cannot be blindly performed in high-dimensions because you will get zero training error but awful test error;
  - Properly estimating the test error is crucial; and
  - There are a few tricks to extend classical classification approaches to high-dimensions, which we have already seen in the regression context!

Classification

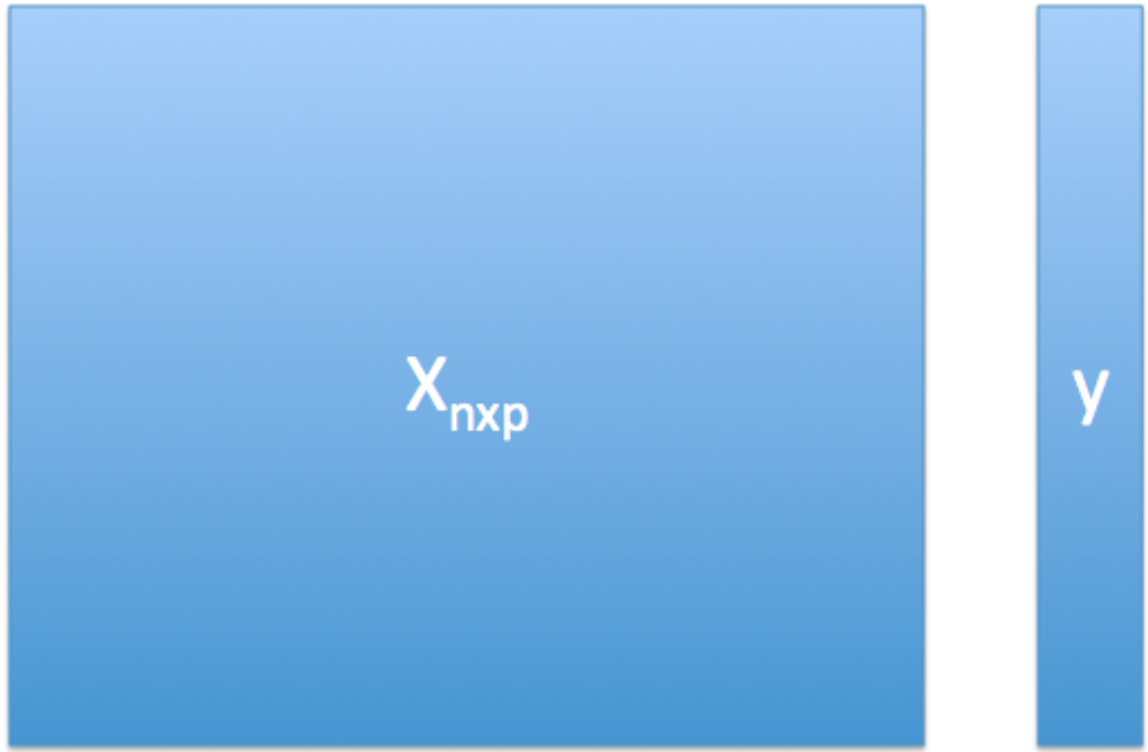
- There are many approaches out there for performing classification.
- We will discuss a few ideas
  - Model-based methods: logistic regression, LDA, QDA.
  - Non-parametric methods: KNN classification.
  - Margin-based classifiers: support vector machines (SVM).

The Classification Task

- Similar to regression, the classification problem is supervised learning:

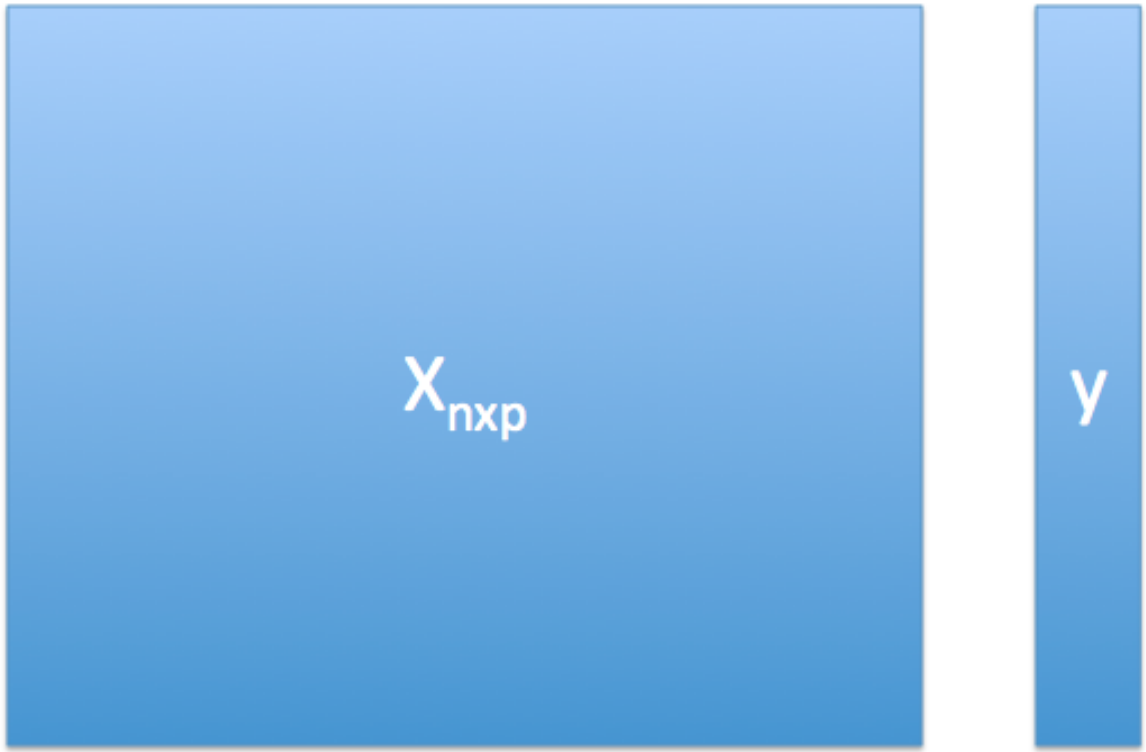
The Classification Task

- Similar to regression, the classification problem is supervised learning:



The Classification Task

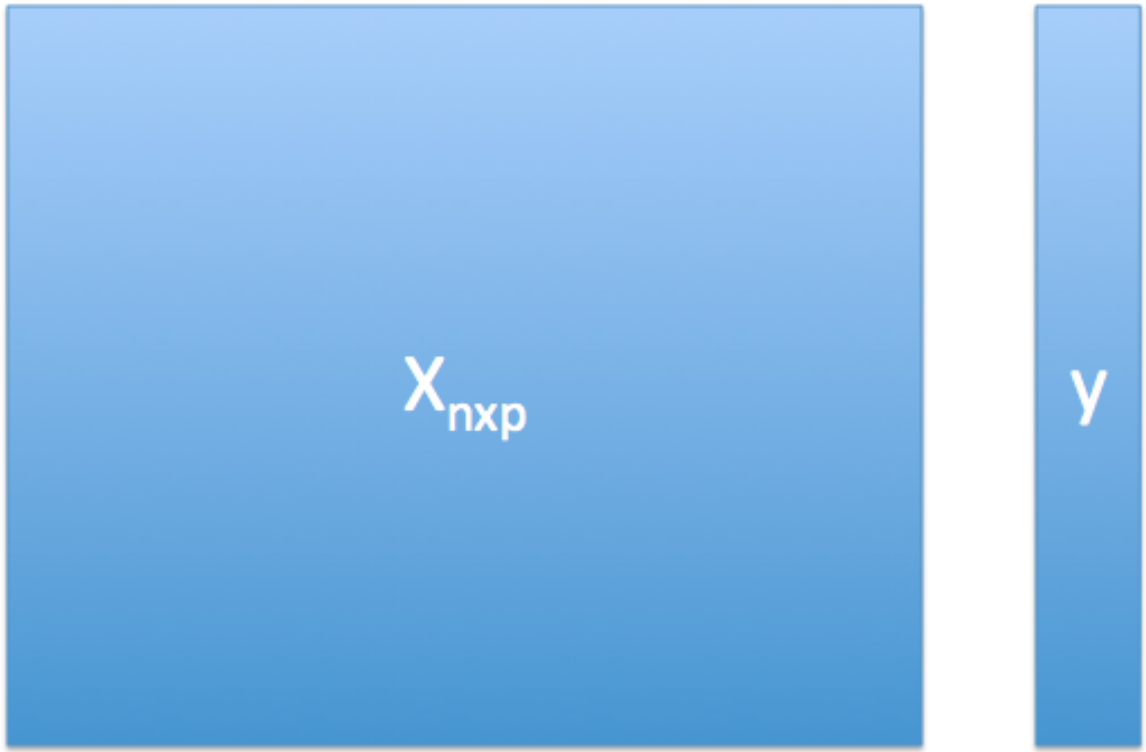
- Similar to regression, the classification problem is supervised learning:



- The only difference is that the response y is a categorical variable with (in general) K categories

The Classification Task

- Similar to regression, the classification problem is supervised learning:



- The only difference is that the response y is a categorical variable with (in general) K categories
- We mostly focus on the case of $K = 2$, i.e. cancer vs benign, but the ideas are the same

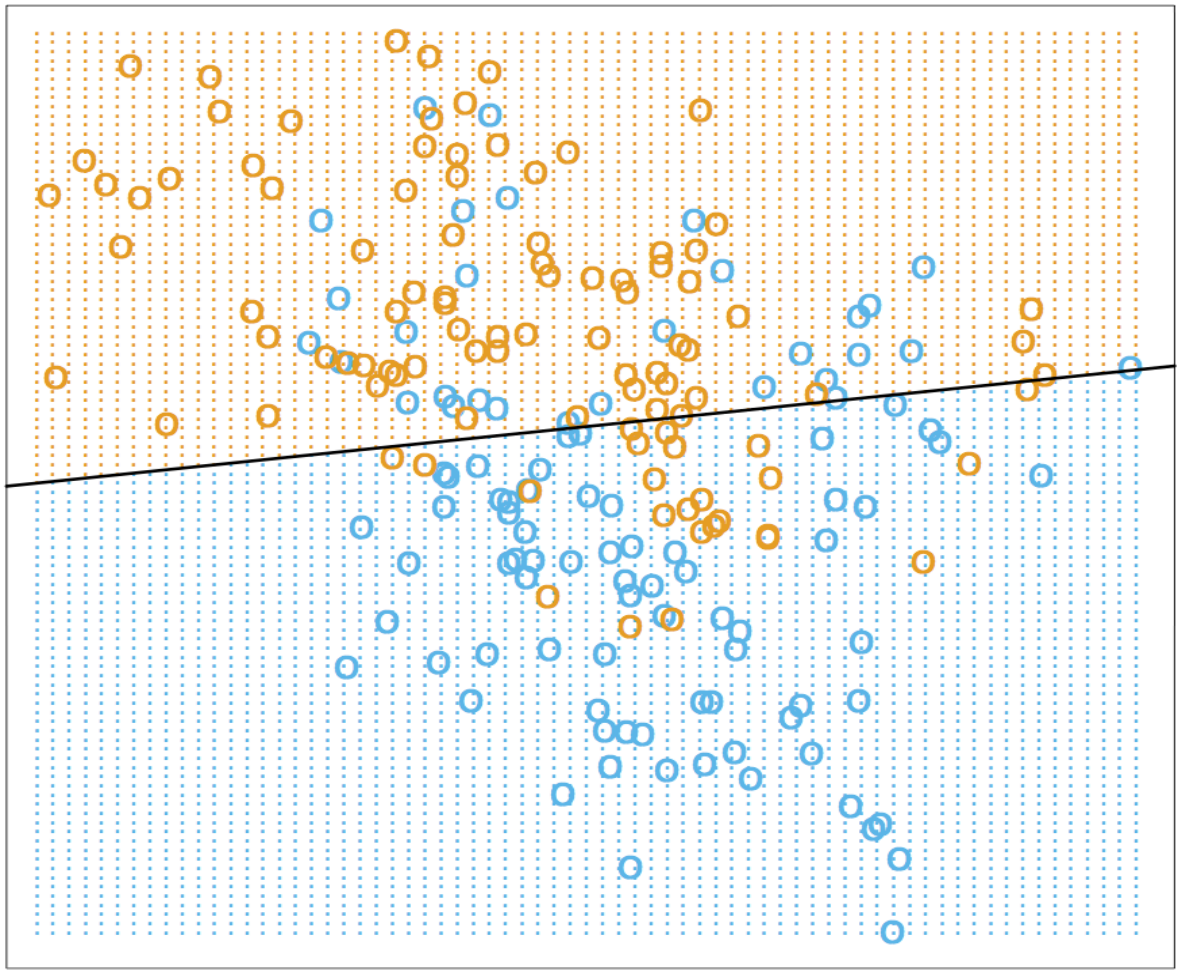
Classification using Linear Regression

Classification using Linear Regression

- There is really nothing preventing us from doing this: we can fit a linear regression with a categorical response!

Classification using Linear Regression

- There is really nothing preventing us from doing this: we can fit a linear regression with a categorical response!



Classification using Linear Regression

Classification using Linear Regression

- Consider the simple case of $p = 1$

Classification using Linear Regression

- Consider the simple case of $p = 1$
- Suppose that y_i can be 1 or -1 (positive or negative) with equal probability

Classification using Linear Regression

- Consider the simple case of $p = 1$
- Suppose that y_i can be 1 or -1 (positive or negative) with equal probability
- In this case, no intercept is needed ($\bar{y} = 0$) so the linear regression tries to find β that minimizes the RSS:

$$\|y - x\beta\|_2^2 = \sum_i (y_i - \beta x_i)^2$$

Classification using Linear Regression

- Consider the simple case of $p = 1$
- Suppose that y_i can be 1 or -1 (positive or negative) with equal probability
- In this case, no intercept is needed ($\bar{y} = 0$) so the linear regression tries to find β that minimizes the RSS:

$$\|y - x\beta\|_2^2 = \sum_i (y_i - \beta x_i)^2$$

- As we discussed before,

$$x_i\hat{\beta} = \hat{y}_i$$

Classification using Linear Regression

- Consider the simple case of $p = 1$
- Suppose that y_i can be 1 or -1 (positive or negative) with equal probability
- In this case, no intercept is needed ($\bar{y} = 0$) so the linear regression tries to find β that minimizes the RSS:

$$\|y - x\beta\|_2^2 = \sum_i (y_i - \beta x_i)^2$$

- As we discussed before,

$$x_i\hat{\beta} = \hat{y}_i$$

- In this case, we set

$$C_i = \begin{cases} 1 & x_i\hat{\beta} > 0 \\ -1 & x_i\hat{\beta} \leq 0 \end{cases}$$

Classification using Linear Regression

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!
- Suppose $y_i = 1$
  - Suppose $\hat{y}_i = 0.1$; then $(y_i - \hat{y}_i) = 0.9$

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!
- Suppose $y_i = 1$
  - Suppose $\hat{y}_i = 0.1$; then $(y_i - \hat{y}_i) = 0.9$
  - On the other hand, if $\hat{y}_i = -0.1$, $(y_i - \hat{y}_i) = 1.1$

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!
- Suppose $y_i = 1$
  - Suppose $\hat{y}_i = 0.1$; then $(y_i - \hat{y}_i) = 0.9$
  - On the other hand, if $\hat{y}_i = -0.1$, $(y_i - \hat{y}_i) = 1.1$

  These are not very different!

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!
- Suppose $y_i = 1$
  - Suppose $\hat{y}_i = 0.1$; then $(y_i - \hat{y}_i) = 0.9$
  - On the other hand, if $\hat{y}_i = -0.1$, $(y_i - \hat{y}_i) = 1.1$

  These are not very different!
- However, in the first case, $C_i = 1$ and in the second case, $C_i = -1$

Classification using Linear Regression

- We can also do this with multiple predictors $p > 1$, and can map any value of y to $\{-1, 1\}$
- This model assumes that the two classes can be separated with a line (hyperplane for $p > 1$), which is somewhat unrealistic!
- Suppose $y_i = 1$
  - Suppose $\hat{y}_i = 0.1$; then $(y_i - \hat{y}_i) = 0.9$
  - On the other hand, if $\hat{y}_i = -0.1$, $(y_i - \hat{y}_i) = 1.1$

  These are not very different!
- However, in the first case, $C_i = 1$ and in the second case, $C_i = -1$
- This suggest that so *sum of squared errors may not be the best loss function for categorical variables*!

Drawbacks of Linear Regression for Classification

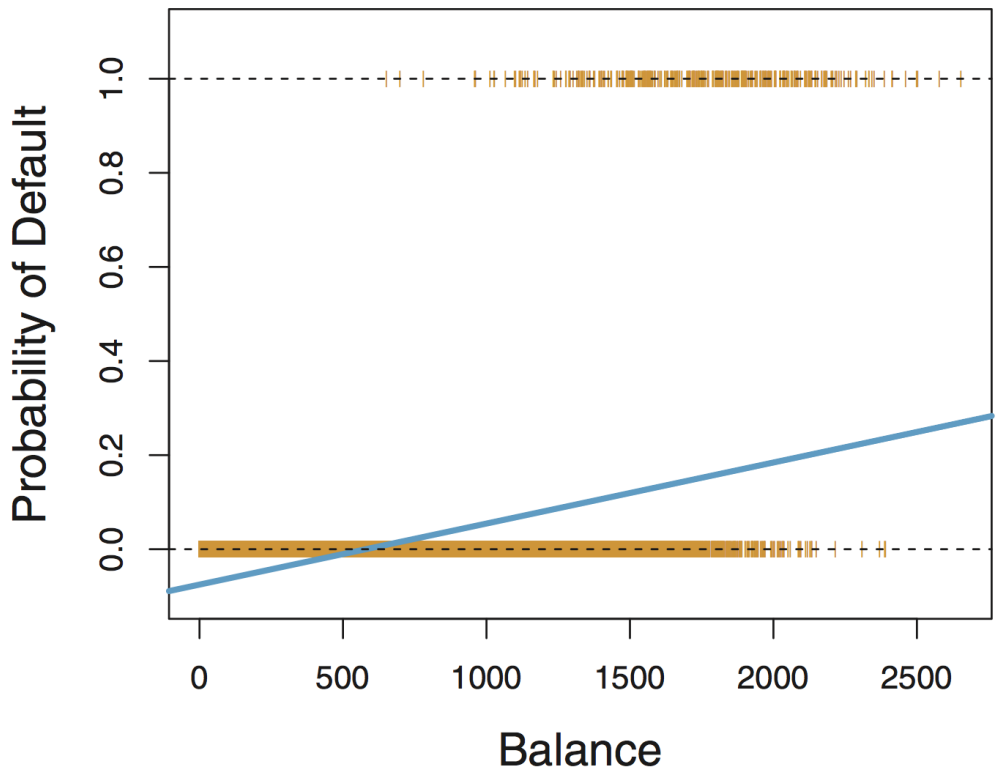
- If we code the values of y as 0 and 1 (instead of -1 and 1), then $X\hat{\beta}$ from linear regression *gives an estimate of the probability* $P(y = 1 \mid X)$, which is sensible.

Drawbacks of Linear Regression for Classification

- If we code the values of y as 0 and 1 (instead of -1 and 1), then $X\hat{\beta}$ from linear regression *gives an estimate of the probability* $P(y = 1 \mid X)$, which is sensible.
- However, there is no guarantee that the estimated probabilities are in fact between 0 and 1!! And, in general, they are actually not!

Drawbacks of Linear Regression for Classification

- If we code the values of y as 0 and 1 (instead of -1 and 1), then $X\hat{\beta}$ from linear regression *gives an estimate of the probability* $P(y = 1 \mid X)$, which is sensible.
- However, there is no guarantee that the estimated probabilities are in fact between 0 and 1!! And, in general, they are actually not!



Drawbacks of Linear Regression for Classification

Drawbacks of Linear Regression for Classification

- There is also a serious problem if y has more than 2 categories!

Drawbacks of Linear Regression for Classification

- There is also a serious problem if y has more than 2 categories!
  - Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms, and there are three possible diagnoses: stroke, drug overdose, and epileptic seizure
  - We could consider a quantitative response as

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

Drawbacks of Linear Regression for Classification

- There is also a serious problem if y has more than 2 categories!
  - Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms, and there are three possible diagnoses: stroke, drug overdose, and epileptic seizure
  - We could consider a quantitative response as

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

  - Unfortunately, this coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure
  - In practice there is no particular reason that this needs to be the case and one could choose any other equally reasonable coding

0-1 Loss and Optimal Classificier

0-1 Loss and Optimal Classificier

- A natural loss function for categorical data is the 0-1 loss function

$$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

  which *counts* how many cases are classified incorrectly.

0-1 Loss and Optimal Classificier

- A natural loss function for categorical data is the 0-1 loss function

$$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

  which *counts* how many cases are classified incorrectly.

- As in the setting of linear regression, in general we want to estimate a function f such that $\hat{f}(x) = \hat{y}$ gives the smallest 0-1 loss.

0-1 Loss and Optimal Classificier

- A natural loss function for categorical data is the 0-1 loss function

$$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

  which *counts* how many cases are classified incorrectly.
- As in the setting of linear regression, in general we want to estimate a function f such that $\hat{f}(x) = \hat{y}$ gives the smallest 0-1 loss.
- Using this loss function, a good classifier is one for which the test error

$$E(y_0 \neq \hat{y}_0)$$

  is minimized.

0-1 Loss and Optimal Classificier

0-1 Loss and Optimal Classificier

- It turns out that the test error based on 0-1 loss is minimized by if we assign each observation to the most likely class, *given its predictor values*

$$P(Y = k \mid X = x_0)$$

0-1 Loss and Optimal Classificier

- It turns out that the test error based on 0-1 loss is minimized by if we assign each observation to the most likely class, *given its predictor values*

$$P(Y = k \mid X = x_0)$$

- This is called the Bayes classifier (the above formula comes from the application of *Bayes Theorem*)

0-1 Loss and Optimal Classificier

- It turns out that the test error based on 0-1 loss is minimized by if we assign each observation to the most likely class, *given its predictor values*

$$P(Y = k \mid X = x_0)$$

- This is called the Bayes classifier (the above formula comes from the application of *Bayes Theorem*)
- Note that in general, we can't use this rule, *unless we know the joint probability distribution*

0-1 Loss and Optimal Classificier

- It turns out that the test error based on 0-1 loss is minimized by if we assign each observation to the most likely class, *given its predictor values*

$$P(Y = k \mid X = x_0)$$

- This is called the Bayes classifier (the above formula comes from the application of *Bayes Theorem*)
- Note that in general, we can't use this rule, *unless we know the joint probability distribution*
- However, the Bayes classifier gives us a "benchmark" in simulation settings, as the beset possible classifier

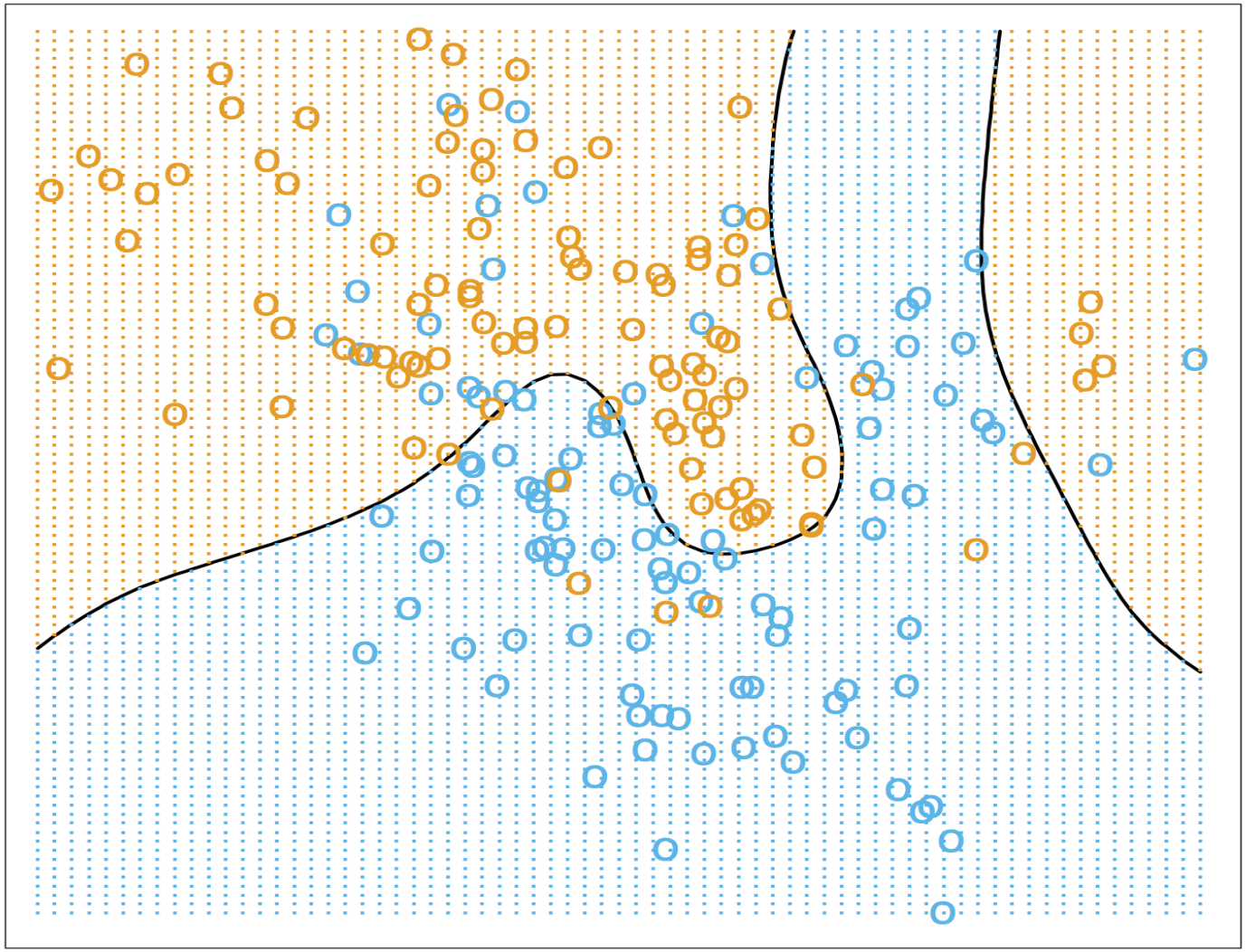
0-1 Loss and Optimal Classificier

- It turns out that the test error based on 0-1 loss is minimized by if we assign each observation to the most likely class, *given its predictor values*

$$P(Y = k \mid X = x_0)$$

- This is called the Bayes classifier (the above formula comes from the application of *Bayes Theorem*)
- Note that in general, we can't use this rule, *unless we know the joint probability distribution*
- However, the Bayes classifier gives us a "benchmark" in simulation settings, as the beset possible classifier
- We can also get the Bayes error rate, which is the *lowest test error we can get*

0-1 Loss and Optimal Classificier



Logistic Regression

- Logistic regression is the straightforward extension of linear regression to the classification setting.

Logistic Regression

- Logistic regression is the straightforward extension of linear regression to the classification setting.
- For simplicity, suppose $y \in \{0, 1\}$: a two-class classification problem.

Logistic Regression

- Logistic regression is the straightforward extension of linear regression to the classification setting.
- For simplicity, suppose $y \in \{0, 1\}$: a two-class classification problem.
- Instead, logistic regression assumes a parametric model

$$P(y = 1 \mid X) = \frac{\exp(X^T \beta)}{1 + \exp(X^T \beta)}.$$

Logistic Regression

- Logistic regression is the straightforward extension of linear regression to the classification setting.
- For simplicity, suppose $y \in \{0, 1\}$: a two-class classification problem.
- Instead, logistic regression assumes a parametric model

$$P(y = 1 \mid X) = \frac{\exp(X^T \beta)}{1 + \exp(X^T \beta)}.$$

- *If this assumption holds*, logistic regression, is a good model-based alternative to Bayes classifier.

Logistic Regression

Logistic Regression

- Taking log and doing some algebra, we can see that

$$\log\left(\frac{P(y=1 \mid X)}{1-P(y=1 \mid X)}\right) = X^T\beta$$

Logistic Regression

- Taking log and doing some algebra, we can see that

$$\log\left(\frac{P(y=1 \mid X)}{1-P(y=1 \mid X)}\right) = X^T\beta$$

- $\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right) = \log\left(\frac{P(y=1|X)}{P(y=0|X)}\right)$ is the log-odds, or logit transform

Logistic Regression

- Taking log and doing some algebra, we can see that

$$\log\left(\frac{P(y=1 \mid X)}{1-P(y=1 \mid X)}\right) = X^T\beta$$

- $\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right) = \log\left(\frac{P(y=1|X)}{P(y=0|X)}\right)$ is the log-odds, or logit transform
- This means that logistic regression is a *linear model* in the new, transformed domain. These types of models are called generalized linear models, and we will see more examples of these later.

Logistic Regression

- Taking log and doing some algebra, we can see that

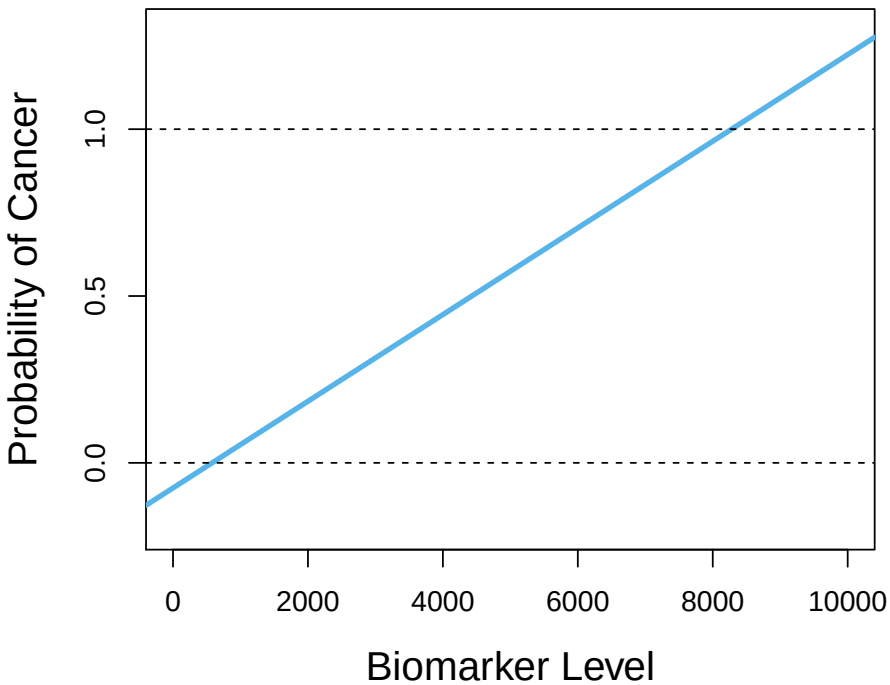
$$\log\left(\frac{P(y=1 \mid X)}{1-P(y=1 \mid X)}\right) = X^T\beta$$

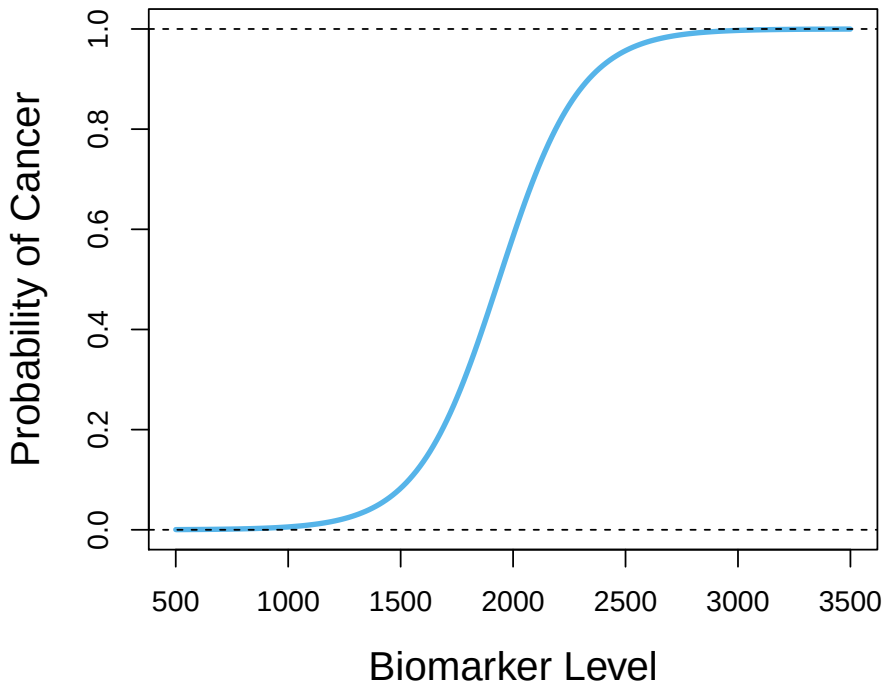
- $\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right) = \log\left(\frac{P(y=1|X)}{P(y=0|X)}\right)$ is the log-odds, or logit transform
- This means that logistic regression is a *linear model* in the new, transformed domain. These types of models are called generalized linear models, and we will see more examples of these later.
- We usually fit this model using maximum likelihood – like least squares, but for logistic regression.

Example in R

```
xtr <- matrix(rnorm(1000*20),ncol=20)
beta <- c(rep(1,10),rep(0,10))
ytr <- 1*((xtr%*%beta + .2*rnorm(1000)) >= 0)
mod <- glm(ytr~xtr,family="binomial")
print(summary(mod))
```

Logistic vs Linear Regression





- Left: linear regression.
- Right: logistic regression.

Five Ways to Extend Logistic to High Dimensions

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection
2. Forward Stepwise Logistic Regression

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection
2. Forward Stepwise Logistic Regression
3. Ridge Logistic Regression

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection
2. Forward Stepwise Logistic Regression
3. Ridge Logistic Regression
4. Lasso Logistic Regression

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection
2. Forward Stepwise Logistic Regression
3. Ridge Logistic Regression
4. Lasso Logistic Regression
5. Principal Components Logistic Regression

Five Ways to Extend Logistic to High Dimensions

1. Variable Pre-Selection
2. Forward Stepwise Logistic Regression
3. Ridge Logistic Regression
4. Lasso Logistic Regression
5. Principal Components Logistic Regression

How to decide which approach is best, and which tuning parameter value to use for each approach? Cross-validation or validation set approach.

Example in R: Lasso Logistic Regression

```
xtr <- matrix(rnorm(1000*20),ncol=20)
beta <- c(rep(1,5),rep(0,15))
ytr <- 1*((xtr%*%beta + .5*rnorm(1000)) >= 0)
cv.out <- cv.glmnet(xtr, ytr, family="binomial", alpha=1)
plot(cv.out)
```

Batch Effects

Batch Effects

- In any sort of omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...
- Similar issues exist in other biomedical big data (e.g. EHR, etc)

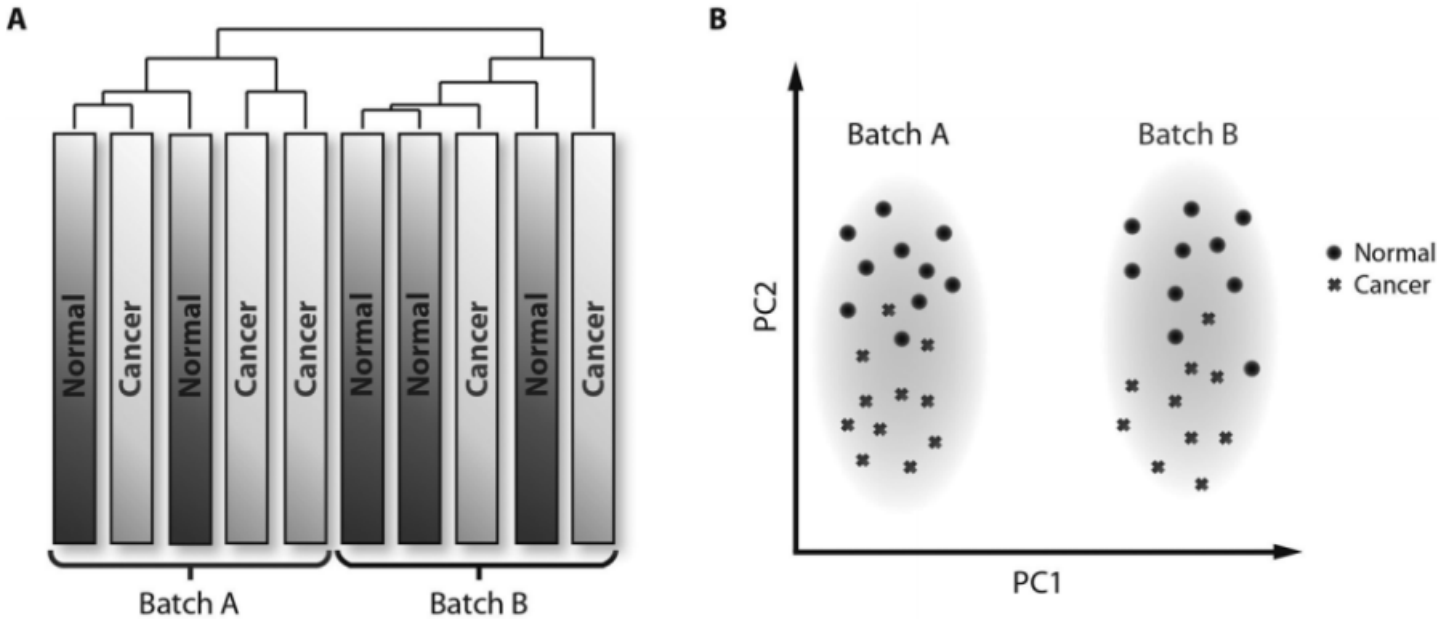
Batch Effects

- In any sort of omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...
- Similar issues exist in other biomedical big data (e.g. EHR, etc)
- It has been shown many many times that batch effects can be much stronger than biological effects of interest!

Batch Effects

- In any sort of omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...
- Similar issues exist in other biomedical big data (e.g. EHR, etc)
- It has been shown many many times that batch effects can be much stronger than biological effects of interest!
- If you are not careful, batch effects can result in complete confounding, making your data worthless and your results nonsense.

Batch Effects



Steps to Reduce Batch Effects

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.
- Need to validate any results obtained on independent data sets from a different institution.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.
- Need to validate any results obtained on independent data sets from a different institution.

Batch effects are almost inevitable. But you can do your best to design an experiment and analyze the data in such a way that batch effects do not compromise the results obtained.

Subtypes of Breast Cancer

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.

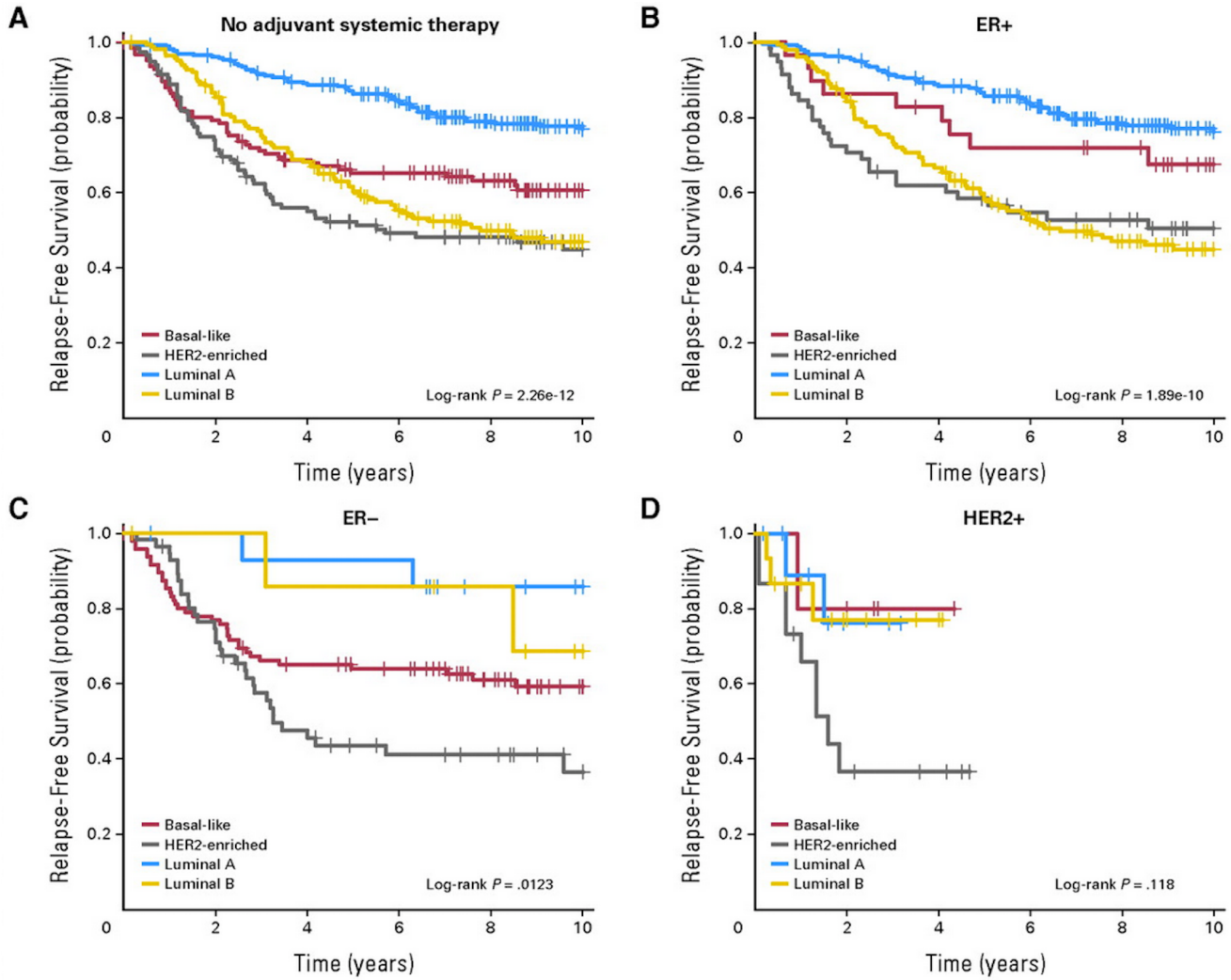
Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.
- Controversy over the best classifier for this task:
  - PAM50 classifier involves 50 genes.
  - More recent proposal involving three genes.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.
- Controversy over the best classifier for this task:
  - PAM50 classifier involves 50 genes.
  - More recent proposal involving three genes.
- Moving target: nobody knows the “true" subtype!
- Prat et al., Breast Cancer Res Treat, 2012

Why Do We Care About Subtypes?



Citation: Parker et al, Journal of Clinical Oncology, 2009

Proteomics for Ovarian Cancer

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.
- Great enthusiasm in the popular press and general public.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.
- Great enthusiasm in the popular press and general public.
- Plans were made to begin marketing a test based on the reported diagnostic.

Not So Fast!!

- Independent researchers took a look at the data, which was publicly available, and discovered:
  - inadvertent changes in protocol mid-experiment: i.e. major batch effects.
  - problems with instrument calibration.
  - difference in processing between tumor and normal samples.

Not So Fast!!

- Independent researchers took a look at the data, which was publicly available, and discovered:
  - inadvertent changes in protocol mid-experiment: i.e. major batch effects.
  - problems with instrument calibration.
  - difference in processing between tumor and normal samples.
- In summary: the observed differences between cancer and normal proteomic patterns were attributable to “artifacts of sample processing, not the underlying biology of cancer."

Gene Expression Signatures for Cancer Treatment

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.
- Several clinical trials were initiated, using these predictors to direct therapy for cancer patients.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.
- Several clinical trials were initiated, using these predictors to direct therapy for cancer patients.
- This research was hailed as a major breakthrough in cancer treatment, and researchers from all over the world tried to use these sorts of techniques in their own labs.

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility
  - Reversal of sensitive/resistant labels

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility
  - Reversal of sensitive/resistant labels
- A shocking paper published by Baggerly and Coombes in Annals of Applied Statistics, detailing all of the errors made: “One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common."

What Went Wrong?

A blasé approach to high-dimensional data analysis:

What Went Wrong?

A blasé approach to high-dimensional data analysis:

- Need to have a proper independent test set, that you simply cannot peek at under any circumstances!

What Went Wrong?

A blasé approach to high-dimensional data analysis:

- Need to have a proper independent test set, that you simply cannot peek at under any circumstances!
- Need to have clearly documented code that contains all steps of the analysis, from start to finish. You must be able to share this code with independent researchers, and you must be confident that your code is correct. If not, then your work isn't ready for prime time.

The Stakes are High!

At Duke:

- Dozens of papers retracted;
- Careers and reputations ruined;
- Patients endangered through unethical clinical trials.

Plus, a 60 Minutes special feature and an Institute of Medicine Committee!!!

Next Lecture

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- KNN Classifier
- Support Vector Machines (SVM)