Logic Trees and SVMs for Improved Prediction from Complex Data

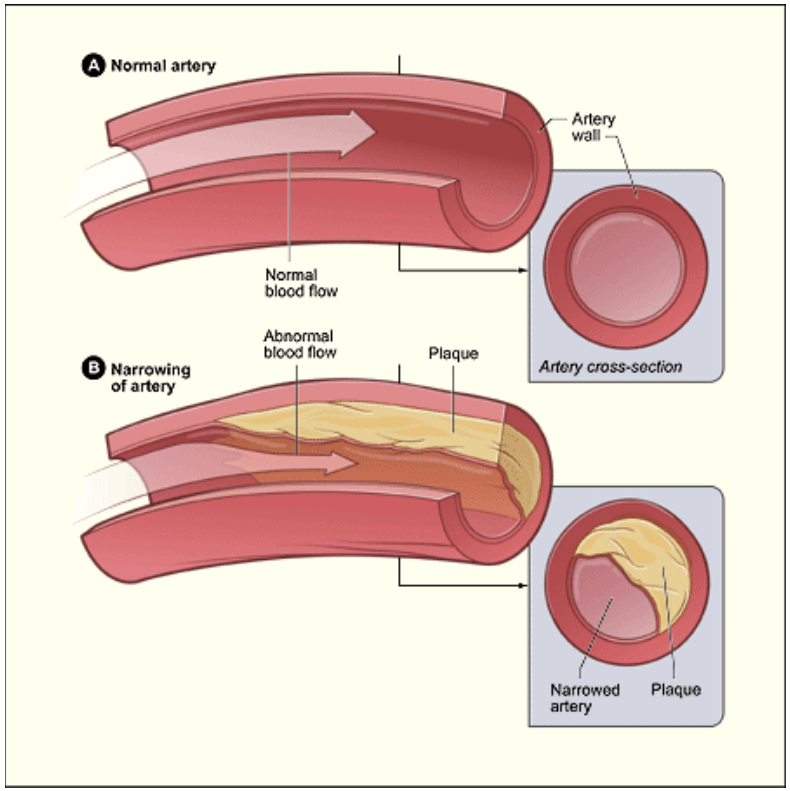
Jennifer Clarke and David Seo

JClarke@med.miami.edu

University of Miami

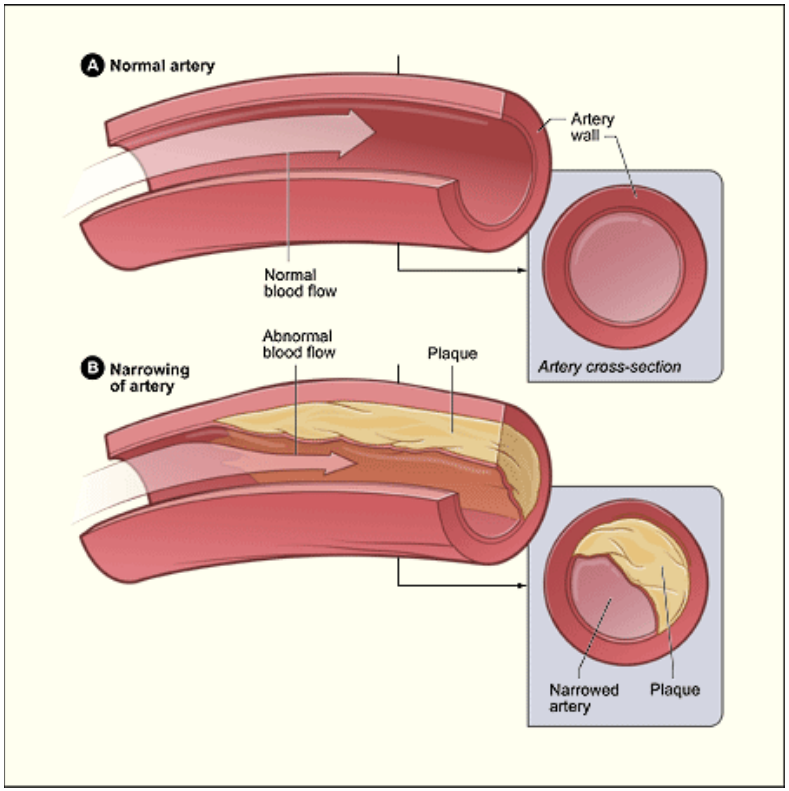
Motivations

- Coronary Heart Disease (CHD) or Coronary Artery Disease (CAD)



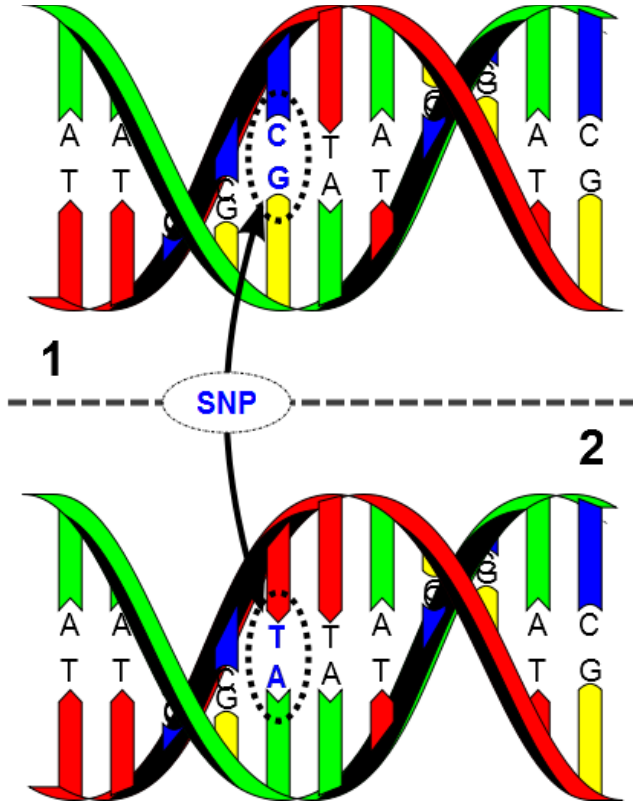
Motivations

- Coronary Heart Disease (CHD) or Coronary Artery Disease (CAD)
- Framingham Coronary Prediction Algorithm



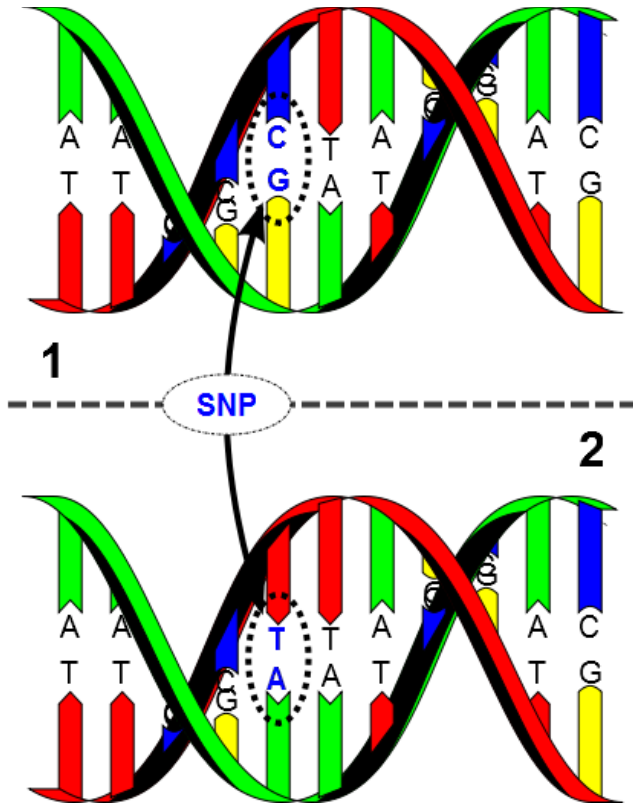
Motivations

- Coronary Heart Disease (CHD) or Coronary Artery Disease (CAD)
- Framingham Coronary Prediction Algorithm
- Single Nucleotide Polymorphism (SNP) Data



Motivations

- Coronary Heart Disease (CHD) or Coronary Artery Disease (CAD)
- Framingham Coronary Prediction Algorithm
- Single Nucleotide Polymorphism (SNP) Data
- Heterogeneous Population: clinical need for predictive *accuracy*



Outline

- Existing Model
- Strategies for Improved Prediction
  - Weighted Average of Predictions
  - Combined Data Model
  - Two-Stage Approach with SVMs
- Model Types
  - Logic Regression
  - Logistic Regression with LASSO
- Support Vector Machines
- Application to CATHGEN population
- Conclusion and Discussion

Framingham Coronary Prediction Algorithm

[Wilson, 1998]

- Estimate of total CHD risk over the course of 10 years
- Separate logistic regression models for men and women
- Age, blood cholesterol, HDL cholesterol, blood pressure, cigarette smoking, diabetes mellitus

Framingham Coronary Prediction Algorithm

[Wilson, 1998]

- Estimate of total CHD risk over the course of 10 years
- Separate logistic regression models for men and women
- Age, blood cholesterol, HDL cholesterol, blood pressure, cigarette smoking, diabetes mellitus

But:

- Only for persons without known heart disease
- Over-identification of candidates for aggressive interventions among elderly
- Does not predict disease burden

Incorporation of Genomic Data

Three modeling strategies to predict $\mathbf{Y}$ (CHD burden) from existing $\mathbf{Z}$ and new $\mathbf{X}$

- *Weighted Average of Clinical Model and Genomic Model*
  $\bar{\hat{\mathbf{Y}}} = \alpha\hat{\mathbf{Y}}_e(\mathbf{Z}) + (1-\alpha)\hat{\mathbf{Y}}_g(\mathbf{X})$
  Utilize existing model predictions globally
- *Composite Model*
  $\hat{\mathbf{Y}}_c = f(\mathbf{Z}, \mathbf{X}) + \epsilon$
  Build new model from scratch
- *Two-Stage Model*
  $\hat{\mathbf{Y}}_s = I_c(f_1(\mathbf{Z})) + I_{\bar{c}}(f_2(\mathbf{X})) + \epsilon$
  Utilize existing model predictions locally via SVM

Model Types

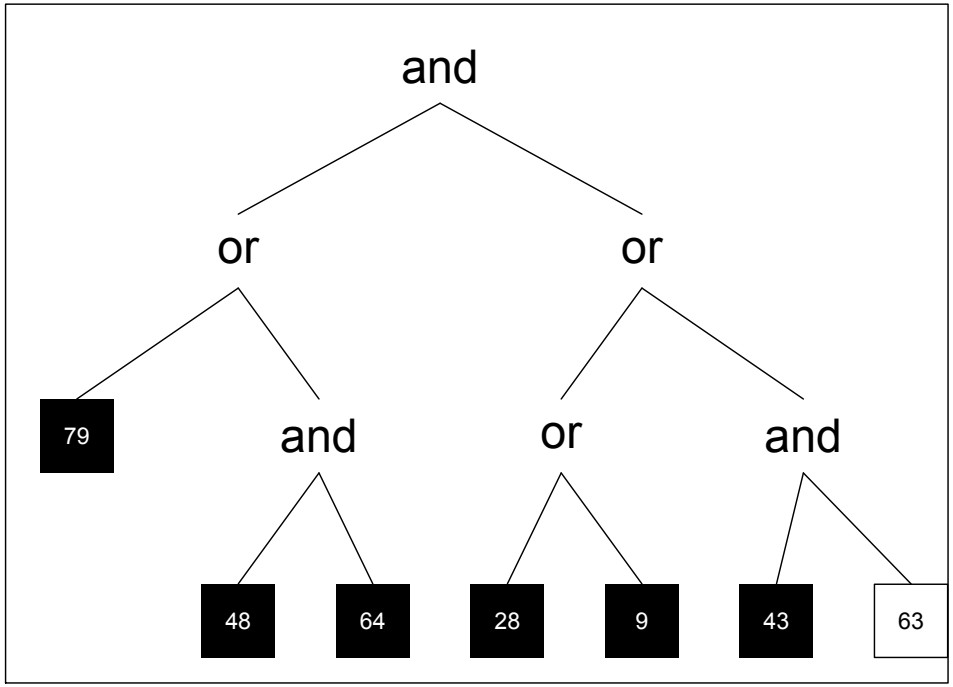
Logic Regression
[Ruczinski, Kooperberg, and LeBlanc 2002]

- adaptive regression for finding Boolean combinations of binary covariates associated with outcome variable
- interaction of many predictors responsible for differences in response
- models of the form $g(E(Y)) = \beta_0 + \sum_{i=1}^{t} \beta_i L_i$ where L_i is a Boolean expression of the covariates and a score function relating the fitted values to the response
  - logistic regression: $g(E(Y)) = \log(E(Y)/(1 - E(Y)))$ with score function RSS
  - classification: $g(E(Y)) = I(L = 1)$ where $I(\cdot)$ is the indicator function

Example

$$(((X_{79}^c) \vee ((X_{48}^c) \wedge (X_{64}^c))) \wedge (((X_{28}^c) \vee (X_9^c)) \vee ((X_{43}^c) \wedge X_{63})))$$

where X_j indicates $X_j = 1$ and X_j^c indicates the conjugate ($X_j = 0$)



Efficient model search via greedy search or simulated annealing

Logic Regression (cont.)

Model Fitting

- training/test set or cross-validation
- randomization testing
  - null model test
  - sequential randomization test (conditional on model size)

Logistic Regression

Variable Selection via LASSO (least absolute shrinkage and selection operator)

[Tibshirani 1996]

$$\min_{\beta_1,\ldots,\beta_m} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 \text{ s.t. } \sum_{j=1}^{m} \mid \beta_j \mid \leq t \text{ for some } t > 0$$

If $t \geq \sum_{j=1}^{m} \mid \hat{\beta}_j^{ols} \mid$ then estimator unchanged. Otherwise LASSO shrinks the estimated coefficient vector towards the origin

LASSO

We can restate the LASSO constraint as a penalized regression: [Osborne, Presnell and Turlach 2000]

$$\min_{\beta_1,\ldots,\beta_m} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{m} \mid \beta_j \mid$$

LASSO

We can restate the LASSO constraint as a penalized regression: [Osborne, Presnell and Turlach 2000]

$$\min_{\beta_1,\ldots,\beta_m} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{m} \mid \beta_j \mid$$

This is similar to *ridge regression* which minimizes:

$$\min_{\beta_1,\ldots,\beta_m} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{m} \beta_j^2$$

Ridge regression scales by a constant factor while LASSO translates by a constant factor, truncating at zero.

Geometry of LASSO

The elliptical contours of $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_j x_{ij})^2$; centered at the OLS estimates; the constraint region.

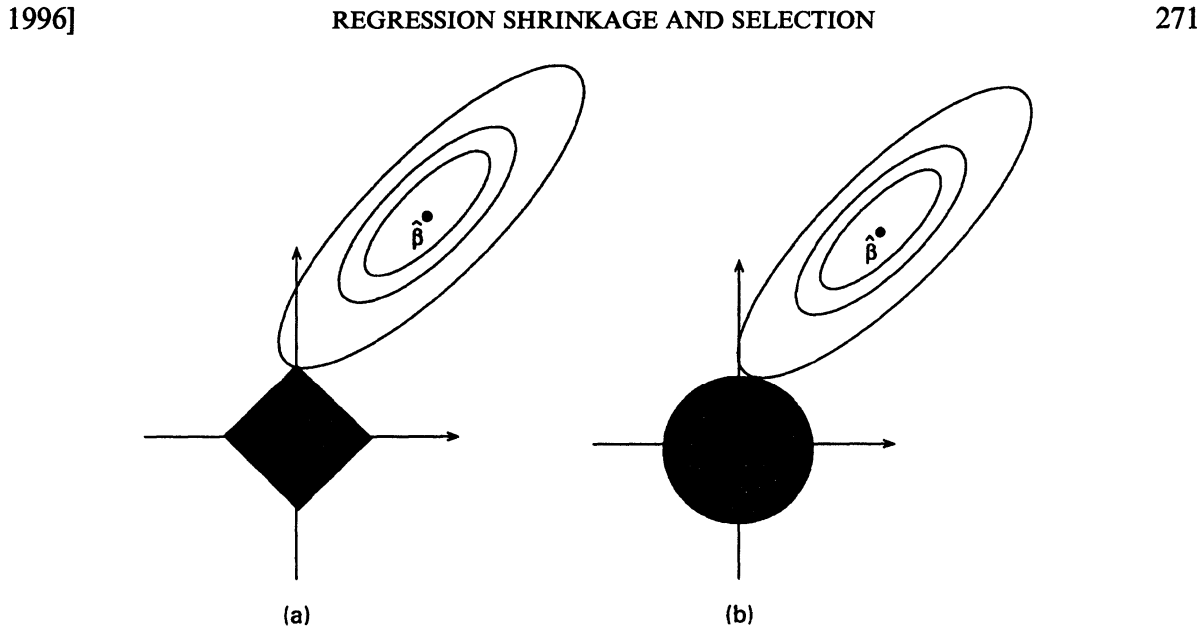


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

The lasso solution is the first place that the contours meet the square; a corner corresponds to a zero coefficient.

Iterative LASSO

- Perform lasso with y as outcome and all available x_i as variables. NOTE: $p \gg n$
- Remove $\{x_k : \hat{\beta}_k = 0\}$. If $\hat{\beta}_j > 0 \; \forall j$, remove $x_k : \hat{\beta}_k = \min\{\hat{\beta}_j\}$.
- Repeat until $p \leq n$.

All remaining variables are used in stepwise logistic regression modeling.

Iterative LASSO

- Perform lasso with y as outcome and all available x_i as variables. NOTE: $p \gg n$
- Remove $\{x_k : \hat{\beta}_k = 0\}$. If $\hat{\beta}_j > 0 \; \forall j$, remove $x_k : \hat{\beta}_k = \min\{\hat{\beta}_j\}$.
- Repeat until $p \leq n$.

All remaining variables are used in stepwise logistic regression modeling.

Is this consistent? Unknown, but similar procedures with lasso have been shown to be consistent
[Wasserman and Roeder 2007, Bunea 2007]

Re: Modeling Strategies

- Weighted Average
- Combined Model
- Two-Stage Approach

Re: Modeling Strategies

- Weighted Average
- Combined Model
- Two-Stage Approach

What is Two-Stage Approach?

S1 Predict y_i using traditional clinical model

S2 If $y_i = \hat{y}_{ie}$, stop. Otherwise, predict y_i using genomic model.

Re: Modeling Strategies

- Weighted Average
- Combined Model
- Two-Stage Approach

What is Two-Stage Approach?

S1 Predict y_i using traditional clinical model

S2 If $y_i = \hat{y}_{ie}$, stop. Otherwise, predict y_i using genomic model.

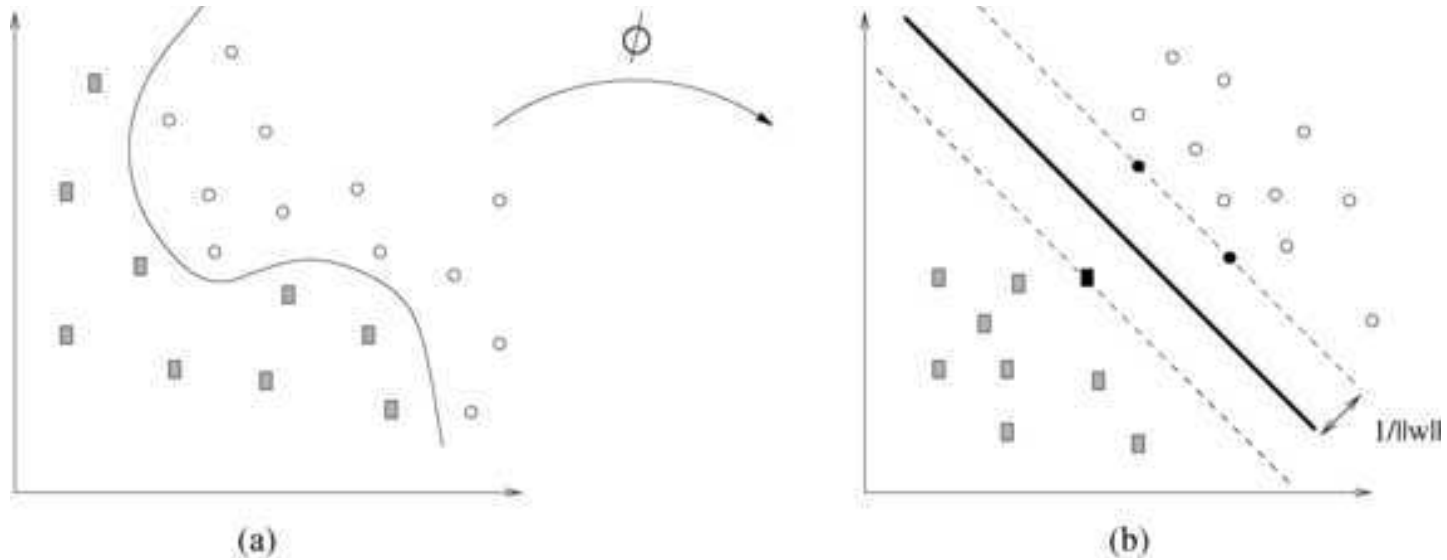
Can we do this without relying on the true value y_i?

YES! Design a support vector machine (SVM) to discriminate those subspaces where the clinical model predicts correctly or incorrectly.

Support Vector Machines (SVMs) for Two-Class Classification

[Boser, Guyon and Vapnik 1992; Moguerza and Munõz 2006]

- Supervised learning methods for classification or regression
- Objective: hyperplane which can separate two classes using covariate information s.t. minimize empirical classification error and maximize geometric margin between classes.
- Estimate linear decision function, where mapping Φ of data from input space into higher-dimensional feature space may be needed



SVMs (cont.)

- Assume Φ exists so data are linearly separable, i.e., $\{(\Phi(x_i), y_i\}_{i=1}^n$ where $y_i \in \{-1, 1\}$.
- Let $w^t\Phi(x) + b = 0$ denote any separating hyperplane in feature space equidistant to nearest point in each outcome class. Rescale w and b:

$$w^t\Phi(x) + b \begin{cases} \geq 1, & \text{if } y_i = +1 \\ \leq -1, & \text{if } y_i = -1 \end{cases}$$

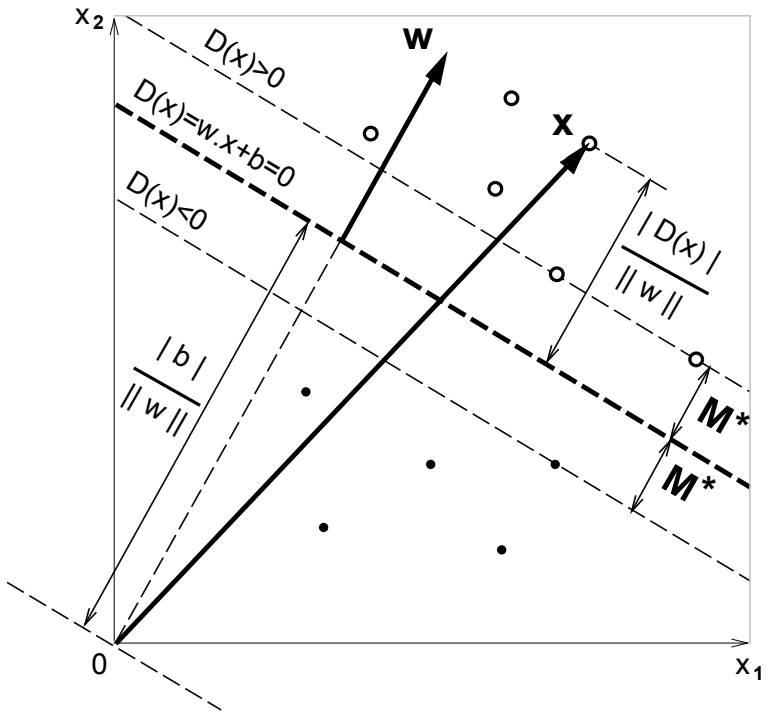
  Note that distance from nearest points to hyperplane is $1/ \parallel w \parallel$.
- To maximize the margin (distance between groups $2/ \parallel w \parallel$) we solve:

$$\min_{w,b} \parallel w \parallel^2 \text{ s. t. } \quad y_i(w^t\Phi(x_i) + b) \geq 1, \ i = 1, \ldots, n.$$

SVMs (cont.)

- Denote solution as w^* and b^*. This determines hyperplane in feature space $D^*(x) = (w^*)^t \Phi(x) + b^* = 0$. The points $\Phi(x)$ that satisfy $y_i((w^*)^t \Phi(x) + b^*) = 1$ are **support vectors**. This is a sparse solution.

Maximum margin linear decision function $D(x) = w \cdot x + b \;\; [\Phi(x) = x]$.



Finding Φ through Kernel Functions

- Φ is hard to find! To avoid explicit knowledge of Φ we use *kernel functions*. A Kernel $K(x, x'),\ K : X \times X \rightarrow \Re$ for which $\exists \Phi$ s.t. $K(x, x') = \Phi(x)^t \Phi(x')$.
- Completion of vector space $f(x) = \sum_j \alpha_j K(x_j, x) + b$ is a reproducing kernel Hilbert space (RKHS). Note $f(x) = w^t \Phi(x) + b$, with $w = \sum_j \alpha_j \Phi(x_j)$, so $f(x) = 0$ describes hyperplane in feature space determined by Φ.
- Since $f(x)$ involves only kernel values, Φ acts implicitly through closed form of K, no explicit evaluation of Φ needed.

Finding Φ through Kernel Functions

- Φ is hard to find! To avoid explicit knowledge of Φ we use *kernel functions*. A Kernel $K(x, x'),\ K : X \times X \to \Re$ for which $\exists \Phi$ s.t. $K(x, x') = \Phi(x)^t \Phi(x')$.
- Completion of vector space $f(x) = \sum_j \alpha_j K(x_j, x) + b$ is a reproducing kernel Hilbert space (RKHS). Note $f(x) = w^t \Phi(x) + b$, with $w = \sum_j \alpha_j \Phi(x_j)$, so $f(x) = 0$ describes hyperplane in feature space determined by Φ.
- Since $f(x)$ involves only kernel values, Φ acts implicitly through closed form of K, no explicit evaluation of Φ needed.

Common kernel functions include

- polynomial $K(x, x') = (c + x^t x')^d$
- radial basis function $K(x, x') = \exp\{-\gamma \mid x - x' \mid^2\}$
- sigmoid $K(x, x') = \tanh\{\gamma x^t x' + c\}$

Analysis Recap

Three modeling strategies:

- Weighted Average
- Combined Model
- Two-Stage Approach
  - SVM will predict whether y_i can be predicted correctly by existing clinical model
  - If yes, use existing model to predict y_i. Else use genomic model to predict y_i

Analysis Recap

Three modeling strategies:

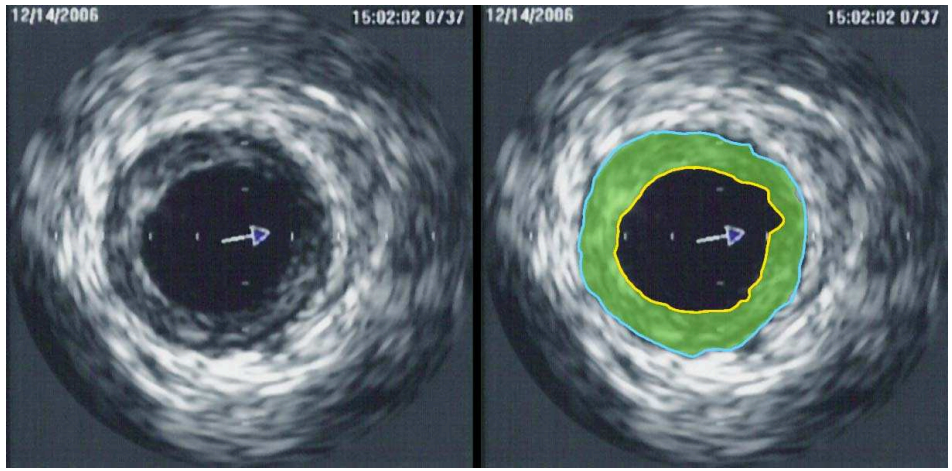
- Weighted Average
- Combined Model
- Two-Stage Approach
  - SVM will predict whether y_i can be predicted correctly by existing clinical model
  - If yes, use existing model to predict y_i. Else use genomic model to predict y_i

Two model types:

- logistic regression
- logic regression
  - logic classification
  - logic logistic regression

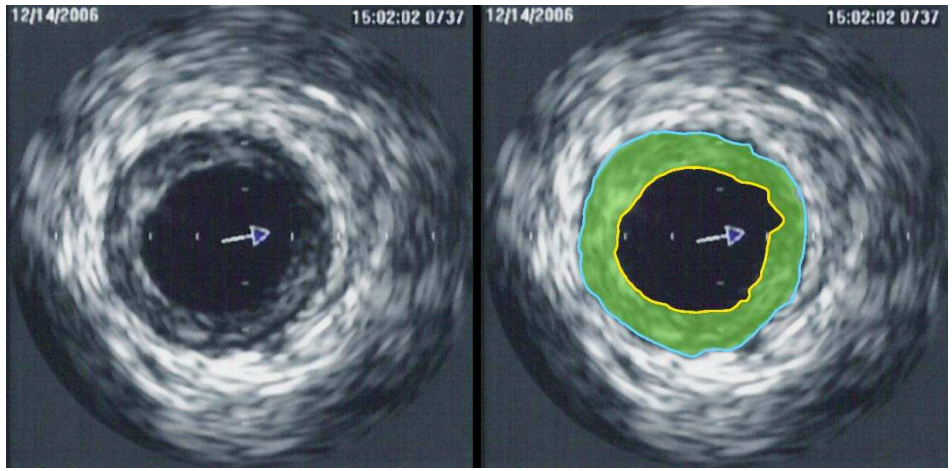
CATHGEN study

- substantial problem in clinical cardiology is gap in ability to detect asymptomatic individuals at high risk for (CHD)



CATHGEN study

- substantial problem in clinical cardiology is gap in ability to detect asymptomatic individuals at high risk for (CHD)



CATHGEN:

- study to evaluate role of genes and gene variants in development of atherosclerosis
- evaluate 1300+ SNPs for association with significant CHD in 1500 subjects who had undergone cardiac catheterization
- 81 SNPs had significant marginal association

CATHGEN study (cont.)

- subset for analysis: 773 white individuals, 385 with SNP data only (264 male, 121 female) and 388 with clinical and SNP data (280 male, 108 female)
- three cohorts within gender
  - (a) controls (age ≥ 65 w/o significant CHD)
  - (b) older cases (age ≥ 65 w/ significant CHD)
  - (c) younger cases (age ≤ 50 w/ significant CHD)
- validation data

(a vs. b) age $56 - 65$ with min or significant CHD

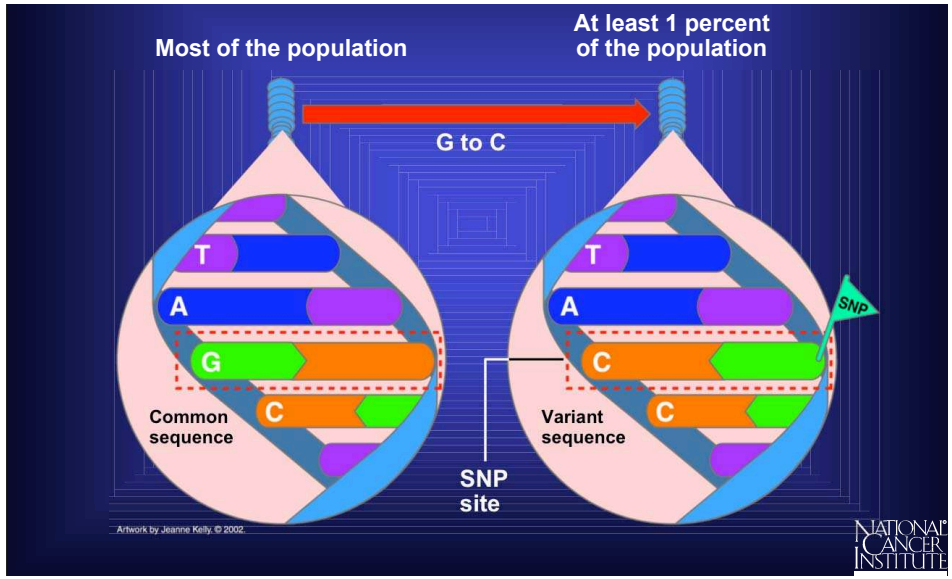
(a vs. c) age $50 - 55$ with min or significant CHD

CATHGEN study (cont.)

		Build Set		Evaluation Set	
		Training	Validation	Training	Validation
Male	Young Cases	44	80	69	103
	Controls	34	14	32	18
	Older Cases	79	13	47	11
Female	Young Cases	11	21	11	18
	Controls	59	12	42	18
	Older Cases	15	3	15	4

Single Nucleotide Polymorphisms (SNPs)

- most common type of genetic variation
- single base change in DNA that occurs in $> 1\%$ of population
- SNPs found in coding and noncoding regions; silent, harmless, harmful, or latent effects
- most frequent *major allele*; less frequent *minor allele*



$$x_i = \begin{cases} 1, & \text{if } \geq 1 \text{ copy minor allele} \\ 0, & \text{otherwise} \end{cases}$$

Model Assessment

- classifying individuals as having non-significant CHD ($Y = 0$) or significant CHD ($Y = 1$)
- fitted or predicted $\hat{Y}_i$ is *accurate* if (1) $Y_i = 0$ and $\hat{Y}_i > 0.5$ or (2) $Y_i = 1$ and $\hat{Y}_i \geq 0.5$ and *inaccurate* otherwise

Model Assessment

- classifying individuals as having non-significant CHD ($Y = 0$) or significant CHD ($Y = 1$)
- fitted or predicted $\hat{Y}_i$ is *accurate* if (1) $Y_i = 0$ and $\hat{Y}_i > 0.5$ or (2) $Y_i = 1$ and $\hat{Y}_i \geq 0.5$ and *inaccurate* otherwise
- model assessment:
  - overall accuracy rate
  - false positive rate ($P(\hat{Y}_i \geq 0.5 | Y_i = 0)$)
  - false negative rate ($P(\hat{Y}_i < 0.5 | Y_i = 1)$)
  - area under the receiver-operating characteristic (ROC) curve (auROC) [probability that model will assign higher $\hat{Y}_i$ to randomly selected CHD sample than to randomly selected control sample]

Computing

- computing in `R`
- `lasso2` [Lokhorst, Venables, Turlach, and Maechler 2006]
- `LogicReg` [Ruczinski, Kooperberg, and LeBlanc 2002]
- `e1071` [Dimitriadou, Hornik, Leisch, Meyer, and Weingessel 2006]
- `ROCR` [Sing, Sander, Beerenwinkel, and Lengauer 2005]

Results of Clinical and Logic Regression Logistic Models on Evaluation Data: Old

	Training/Test Samples				Validation Samples			
Female, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	71.38	52.50	72.50	**84.63**	50.48	59.05	48.57	**79.53**
acc	71.93	49.12	63.16	**84.21**	50.00	54.55	31.82	**72.73**
fn	36.00	20.00	44.00	**12.00**	57.14	28.57	42.86	**14.29**
fp	21.88	75.00	31.25	**18.75**	46.67	53.33	80.00	**26.67**
Male, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	81.97	59.38	82.97	**88.39**	59.52	43.81	63.33	**82.86**
acc	78.48	49.36	82.28	**91.14**	51.72	44.83	51.72	**82.76**
fn	11.48	59.02	**4.92**	6.56	**14.29**	85.71	**14.29**	**14.29**
fp	55.56	22.22	61.11	**16.67**	80.00	26.67	80.00	**20.00**

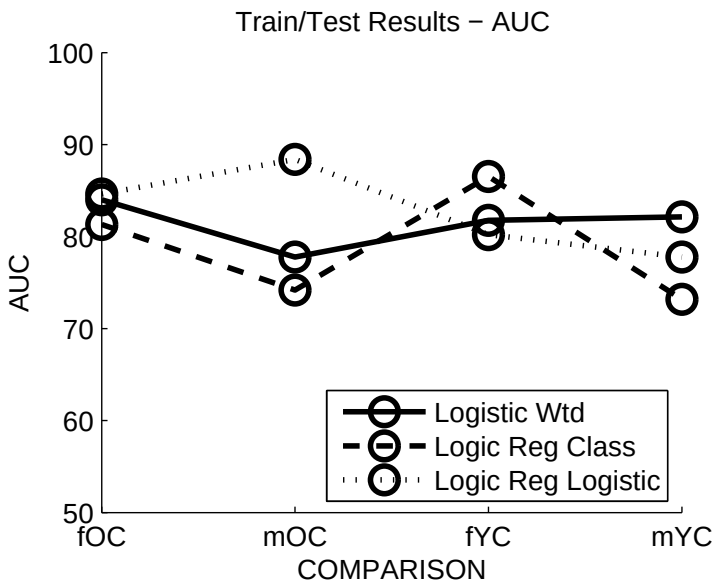
Results of Clinical and Logic Regression Logistic Models on Evaluation Data: Young

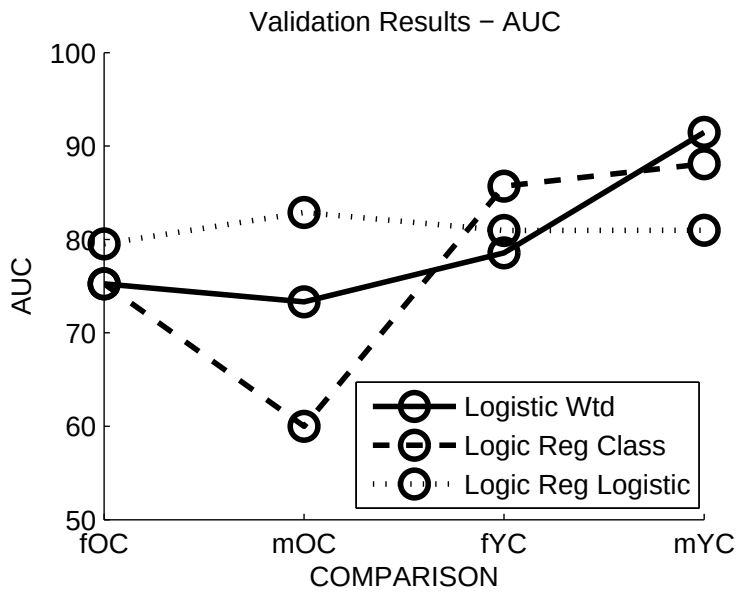
	Training/Test Samples				Validation Samples			
Female, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	80.80	52.60	**81.40**	80.21	63.81	50.00	64.13	**80.95**
acc	79.25	56.60	70.70	**83.02**	61.11	47.22	55.56	**77.78**
fn	**33.33**	66.67	47.62	**33.33**	42.86	66.67	47.62	**38.10**
fp	12.50	28.13	15.63	**6.25**	33.33	33.33	40.00	**0.00**
Male, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_g$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	84.00	47.52	**91.43**	77.78	60.19	57.83	61.32	**80.97**
acc	83.17	49.50	81.13	**92.08**	81.82	57.20	78.51	**91.74**
fn	4.82	49.40	4.82	**0.00**	8.49	44.34	12.26	**4.72**
fp	72.22	55.56	50.00	**44.44**	86.67	40.00	86.67	**33.33**

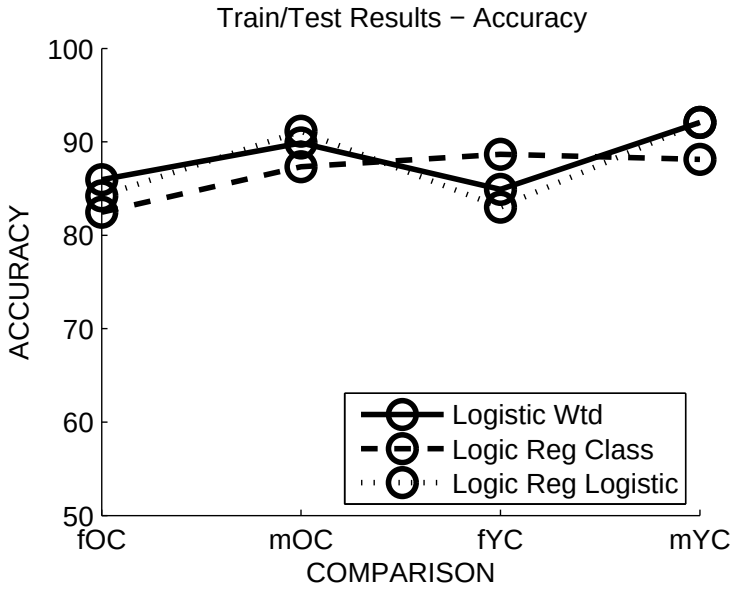
Results (cont.)

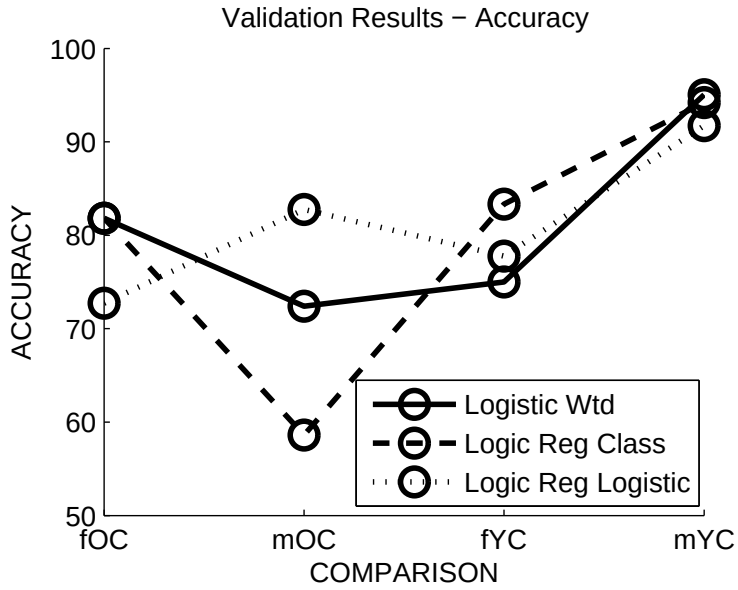
- two-stage approach yields best results
- combined model did not perform better than the clinical only or SNP only models
- weighted average of clinical only and SNP only predictions not competitive; heterogeneity?
- some cohorts have small sample sizes so results considered only ‘proof of concept’
- SNP selection varies with cohort
- no single model type performs consistently best in all comparisons

Results (cont.)









Conclusion

- Two-stage approach to generating predictions from models built from different data sources of multiple types (clinical risk factors, genetic SNP data)
- Performs well for heterogeneous populations for which no single data type is highly informative for all samples
- Validation accuracies are high (between 81.82% and 94.21%) but inferences conditional on fixed chosen model; model uncertainty?
- limited by data and population

Further Research

- other model classes: random forests, neural networks, projection pursuit
- boosting of classifiers?
- variable selection
  - LARS (Efron et al. 2004)
    group-LARS/-LASSO (Yuan and Lin 2006; Park and Hastie 2006)
  - elastic nets (Zou and Hastie 2006)
  - model-free selection (Li, Cook and Nachtsheim 2005)