TOWERS PERRIN

TILLINGHAST

What is Predictive Modeling?

Casualty Actuaries of the Northeast

Spring 2005

Sturbridge, MA

March 23, 2005

Presented by Christopher Monsour, FCAS, MAAA

What is it?

- Estimation of likely outcomes based on historical data
- The emphasis is on estimating the parameters as a means to estimating the outcomes
  - As opposed to financial modeling, where the emphasis is on modeling the probabilities of various outcomes, *given* the parameters
- The emphasis is on different estimates for different combinations of characteristics or for different entities
  - In financial modeling, the emphasis is on the range of possible outcomes for a single entity
- Thus, predictive modeling belongs to statistics and data mining
  - Whereas financial modeling largely belongs to probability theory
- Finally, emphasis on **predictions**, NOT on interpreting model parameters
  - May “interpret” parameters when building model, but only as a means to developing the best model

What sort of outcomes?

Quantitative (regression models)

- The expected length of time to repair an automobile, given
  - Its make, model, and model year
  - The nature of the repair
  - The technician assigned
  - The day of the week service began
- The expected losses for an insured based on that insured's
  - Driving record
  - Age, sex, marital status
  - Location
  - Credit Rating
  - Occupation

What sort of outcomes?

Categorical (classification models)

Soft assignment

- The probability of your home being broken into, depending on
  - Your location
  - The life-stage of your household
  - Whether you have a burglar alarm
  - Whether you have a garage
- The probability that an insured will buy pet health insurance if asked, based on
  - Age, sex, marital status
  - Location
  - Occupation
  - Household type
  - Home value

What sort of outcomes?

Qualitative (classification models)

Hard assignment

- Often soft assignment model plus a threshold, but not always
- Classic example…to which subspecies does a particular botanical specimen belong, based on:
  - Dimensions
  - Coloring
- Is a claim fraudulent?
  - Characteristics of claim
  - Of doctors and lawyers involved
  - Of claimant
  - Of agent or broker

What tools come from predictive models?

Rating factors

- In a linear regression model or a GLM, the model parameters may be interpreted more-or-less directly as indicated rating factors in an additive or multiplicative rating scheme (depending on the type of model)
- The model parameters in a loss ratio model may be interpreted as the amount by which the rating factors need to change

What tools come out of it?

Scoring

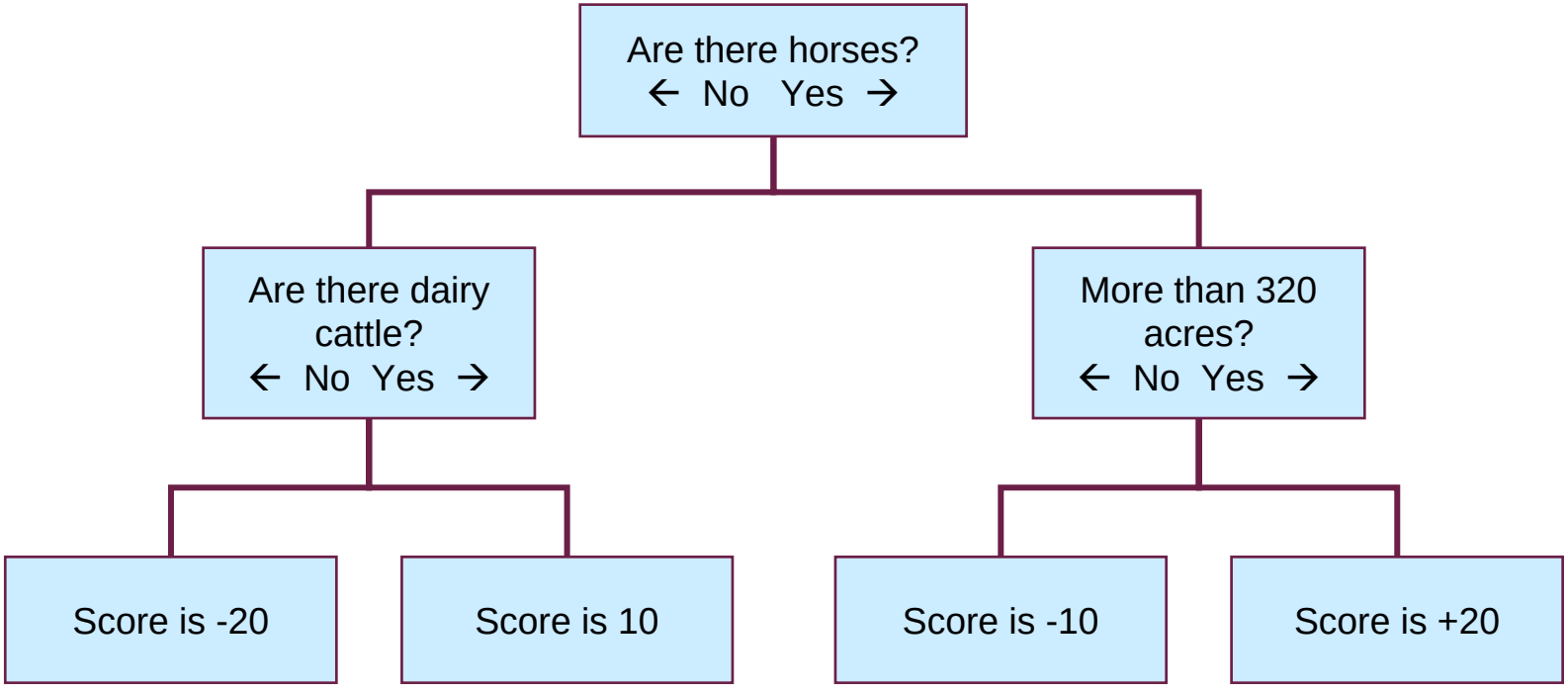
- In a GLM or linear regression, the model scores are added up and a different treatment applied to various ranges of scores, such as
  - Tier assignment for rating / underwriting
  - Adjuster assignment for claims

Farm — Illustration of Scoring	
Animals	
Horses	20
Sheep	-10
Cattle (ranch)	-30
Cattle (dairy)	20
Size of Farm	
< 50 acres	-20
50-100 acres	0
100-320 acres	10
320-640	0
640+	-30
Crops	
Wheat	-30
Barley	10

What tools come out of it?

Rules

- Other models produce branching rules



Related types of modeling

"Unsupervised" learning

- Categorical modeling where the categories are not determined in advance
- Effectively amounts to looking for dense patches, or "clusters", in an appropriate feature space
- Classic example is subspecies classification when name and number of subspecies is unknown in advance
- Geographic use in insurance
  - Feature space can be one dimensional, e.g., pure premium
  - Or can be multi-dimensional, e.g., crime rate, percentage of housing units occupied by owners, etc.

Related types of modeling

Cause-and-effect

- About interpretation of parameters
- Is a certain model of automobile more dangerous than another?
  - Suppose you attempted to answer this just from accident data or insurance data
  - Think about what you might miss
- Well-known that the sign of a coefficient for a predictor can change in a regression model as you add more predictors
  - Is the model correctly specified?
  - Have you added all the predictors you should have out of a possibly infinite number?
- Much more difficult to validate than predictions
- There are specialized methods, used especially in psychology

What do other people do with predictive modeling?

- Pattern recognition / image processing
- Measuring medical trial outcomes
- Direct response modeling
- Classification of texts and artifacts on stylistic and physical criteria
- Categorization of web pages / organization of information
- Planning of product location in stores, to maximize impulse purchases

What good is it in insurance?

- Underwriting / pricing … how to rate, whom to write, what information to pull
  - Claim frequency / claim occurrence model
  - Claim severity models
  - Pure premium and loss ratio models
  - Probability of finding a derogatory if
    - Pull MVR
    - Inspect home
    - Pull arrest record, etc.
- Response models (direct mail, cross-sale) … whom to solicit
- Customer retention
- Premium audit … whom to audit
- Fraud … which claims to refer

Process of predictive modeling

- Lots of vetting of data for unusual values
- Missing values
  - Explore why they are missing
  - Look for correlations with other variables
    - E.g., other variables with missing values
    - Do the missing values come from certain groups?
- Distributions of values
  - Do these change significantly over time?
- Were predictors recorded before the effects you are trying to predict?
  - If not, could the putative predictor be a result of the effect?
- Could something sinister (e.g., a data handling error) explain a powerful model?
- 80% of the time spent on predictive modeling is on this type of work
  - Can't skip, since you **will** end by modeling flukes if you leave them in the data

Topics

- Regression Modeling
- Discrete Modeling
- Model Validation
- Regularization

Regression Models

Linear regression

Least squares estimation

- $y = \beta_0+\beta_1 x_1+\ldots+\beta_n x_n$
- Select the β_i to minimize the sum of squared deviations from the data
- Corresponds to maximum likelihood estimation assuming that y is conditionally normal, given the $x_{i,}$ with variance σ^2 independent of the x_i
- Nice asymptotic properties just given the information about the conditional moments in the above bullet, even if y is not conditionally normal
- Comparison of actual with predicted values of y is used to estimate both the standard error the [normal] distributions from which the values of y are drawn, and the standard errors of the parameter estimates

Linear regression — dummy variables

Categorical predictors

- Effectively, use dummy variables to code all the categories but one:
  - If deductibles are $250, $500, and $1000, with $250 the most common, then have
    - X_1=1 if deductible is $500 and 0 otherwise
    - X_2=1 if deductible is $1000 and 0 otherwise
    - Parameter gives contrast with base class
  - For ordinal categorical variables, sometimes best to code
    - X_1=1 if deductible is >$250
    - X_2=1 if deductible is > $500
    - Parameter gives contrast with adjacent class
    - Now dropping a variable from model = combining adjacent bins

Generalized linear models — exponential family formulation

$E(y) = f(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)$

- f any differentiable monotone function
  - Common choice is the exponential function, which yields a multiplicative model
- y has a conditional distribution (given the x_i) from the "exponential" family (normal, gamma, Poisson, inverse Gaussian, etc.)
- Estimate the β_i by maximum likelihood
- Not the place to give the formulas, but an exponential family has two parameters:
  - Canonical parameter determines the mean, and depends on the x_i
  - Dispersion parameter affects the variance but not the mean, and does not depend on the x_i
    - — Corresponds to the standard error in a linear regression model
    - — But in general, the variance depends on the canonical parameter (or, equivalently, the conditional mean) also
- Significance tests and standard errors of parameter estimates depend on the dispersion parameter, which should not be estimated by maximum likelihood, but (as in the linear regression case) by the sizes of the residuals
  - MLE of the dispersion parameter can be unstable

Generalized linear models — and Heteroscedasticity

- What is the essential difference from linear regression?
  - For continuous y, **not** the functional form of the model
  - $y = f(\beta_0+\beta_1x_1+\ldots+\beta_nx_n)$ can be modeled as a linear regression, just take
    - $f^{-1}(y)= \beta_0+\beta_1x_1+\ldots+\beta_nx_n$
  - Of course, this gives constant variance [or normal distribution] for $f^{-1}(y)$ rather than for y
  - Distributions other than normal allow the variance to be a function of the mean (the predicted value)
  - It turns out that just as a linear regression model can be viewed as minimizing squared-error, without any reference to maximum likelihood, a GLM can be viewed the same way
  - y as a function of x is assumed (at specific values of the x_i) to have mean $\mu=f(b_0+b_1x_1+\ldots+\beta_nx_n)$ and variance $\sigma^2V(\mu)$, where σ^2 is a constant
    - Linear regression corresponds to V constant

Generalized linear models — quasi-likelihood formulation

- So the essential difference is how the variance structure is handled
- Just as for linear regression, we can dispense with distributions
- If we want a specific variance function V(μ), we simply define a quasi-likelihood function, as the sum over all observations of:

$$Q(\mu; y) = \int_{y}^{\mu} \frac{y - t}{\sigma^2 V(\mu)} dt$$

- It's not so easy to maximize that, since we don't know σ^2, but we can maximize the quasi-deviance, which is just $-2\sigma^2 Q$
- So for a GLM, we can define what transform of y will have conditional variance that only depends on its conditional mean, AND we can specify that varoance as a function of the mean

Generalized linear model — typical variance functions

- $V(\mu)=1$ corresponds to least squares (normal distribution)
- $V(\mu)=\mu$ corresponds to the Poisson distribution, which is strictly only applicable to count data, but the quasi-likelihood formulation applies just as well to continuous y
- $V(\mu)=\mu^2$ corresponds to conditional gamma models
- Other quadratic functions of μ correspond to binomial, negative binomial, and hyperbolic secant distributions

Building a model

Continuous predictors

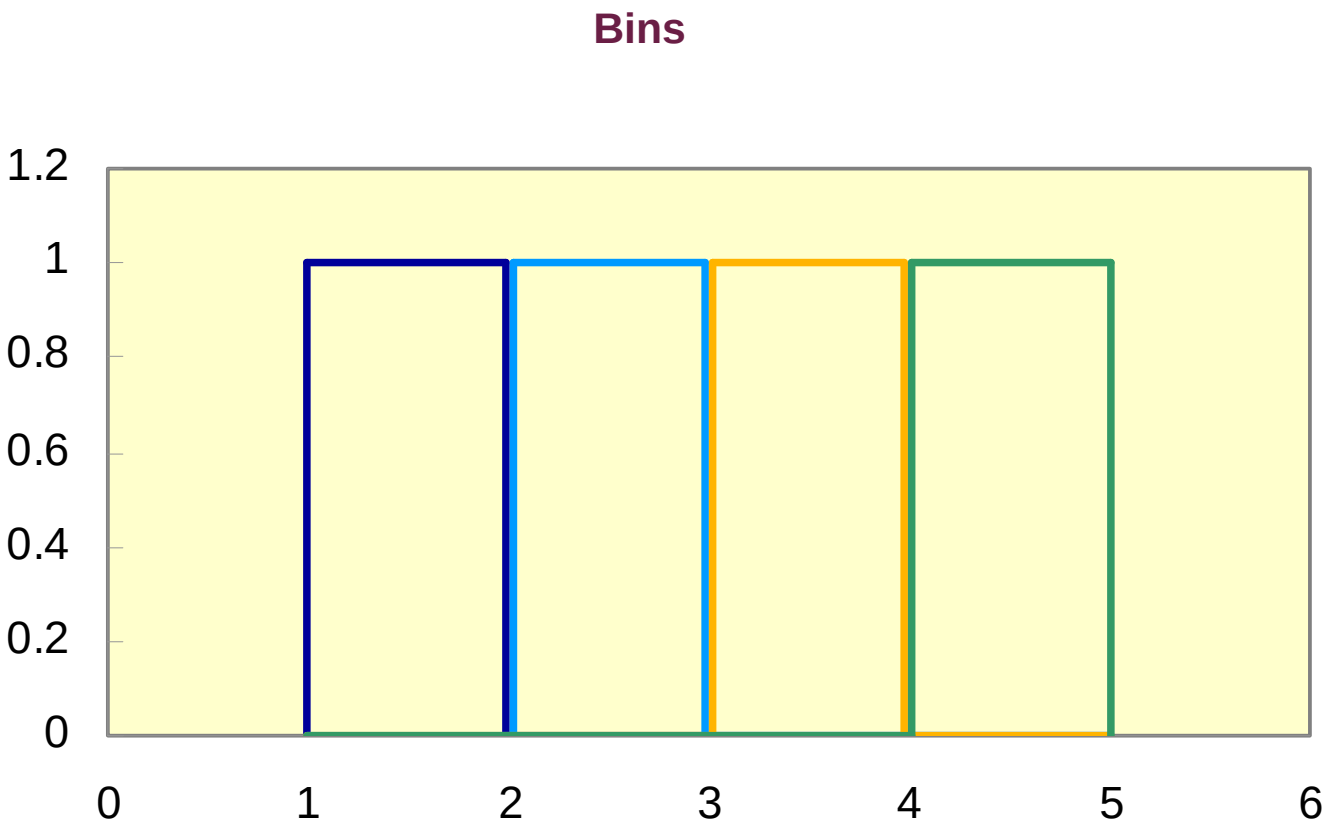
- Often break continuous predictors into ranges and treat as categorical or ordinal….so that one does not need to assume that dependence is linear
- Alternative: Instead of dummy variables like x_i=1 if q > 1000, 0 else, use x_1=q-1000 if q>1000, 0 else
  - Also, functions of the form x_1=1200-q if q<1200, x1=0 otherwise
- Often have a separate dummy to flag whether a specific variable is missing

Interactions

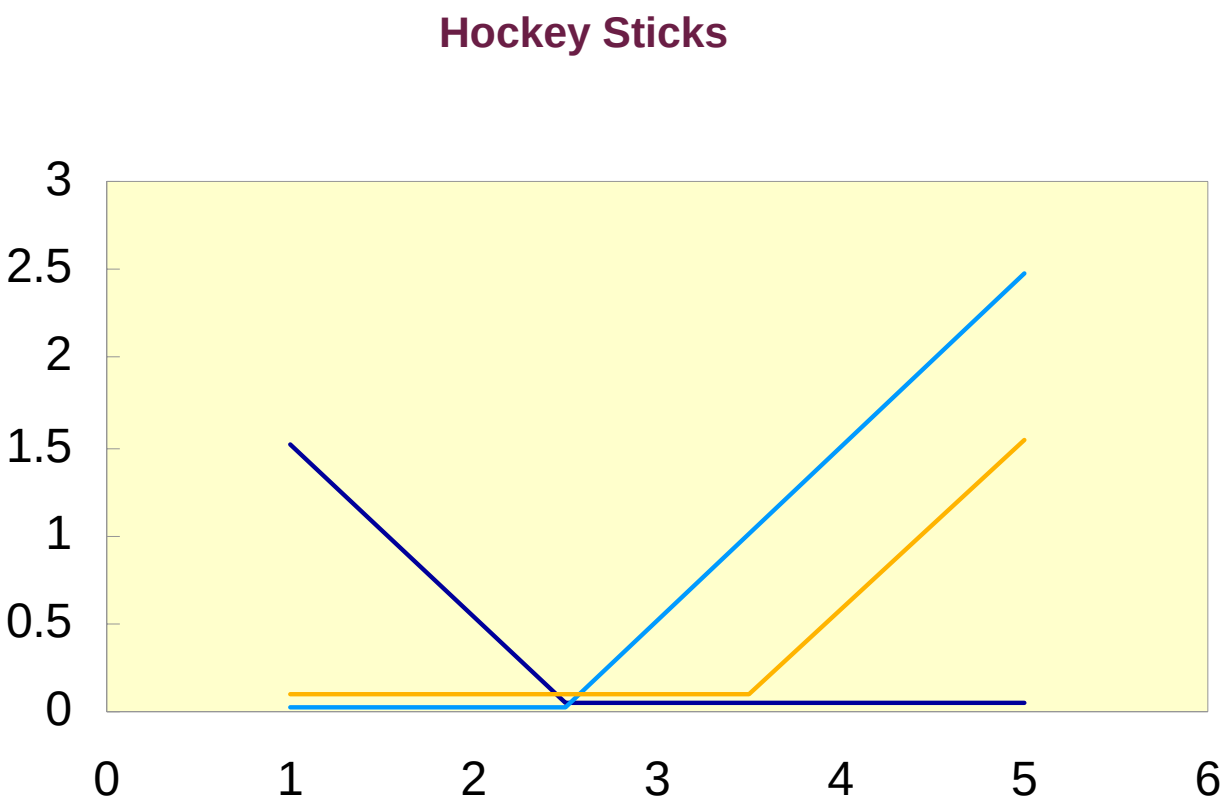
- For predictors treated as continuous, include product x_1x_2 as predictor
- For categorical variables, include dummy variables for all combinations (save base class with base class) of the levels of the two variables
  - Sometimes combine classes, to avoid too many degrees of freedom in the model

	Age			
	18 – 25	26 – 65	66 – 75	76+
Male	x_{m1}	x_{m2}	x_{m3}	x_{m4}
Female	x_{f1}	Base	x_{f3}	x_{f4}

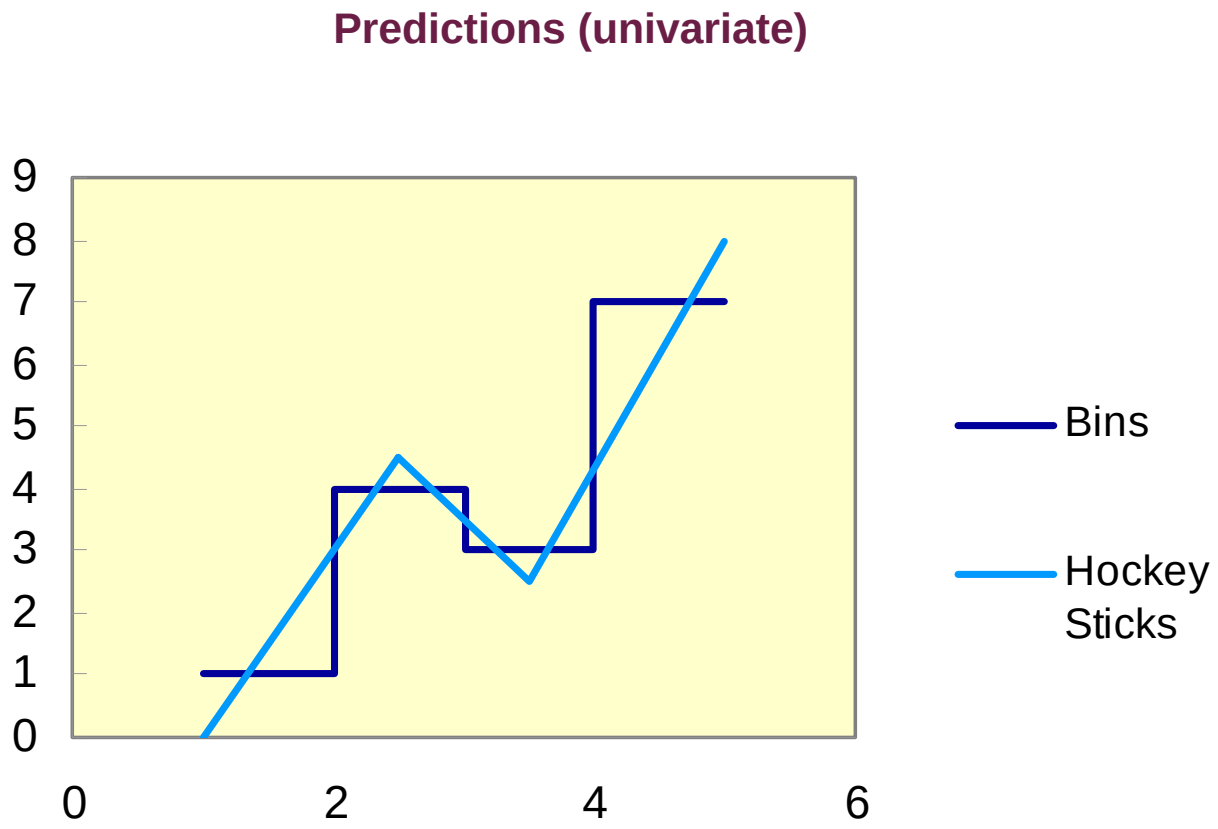
Building a model



Building a model



Building a model — Univariate example of bins vs. hockey sticks



Building a model — stepwise

Automated feature selection

- Stepwise regression (or GLM)
  - Add most significant candidate predictor if significant at a pre-set level
  - Throw out least significant predictor if not significant at a pre-set level

Building a model — MARS, GAM

MARS

- MARS (multivariate adaptive regression splines) searches quite extensively for optimal linear regression models involving the “hockey stick” functions
  - A stepwise regression using these might be called a “poor man’s MARS”

GAM

- Similarly, a generalized additive model (GAM) tries to take non-linearity into account by replacing $\beta_i x_i$ with $g_i(x_i)$, so that

  $y = f(\beta_0 + g_1(x_1) + \ldots + g_n(x_n))$

  where the g_i are cubic smoothing spline functions
  - One can try to fit a spline (not necessarily a cubic smoothing spline) for the most important predictor variable on a univariate basis and then do this recursively for the additional variables
  - “Poor man’s GAM”, or at least “poor man’s additive model”

Discrete Models

Some contrasts

- Supervised vs. unsupervised learning
- Hard vs. soft assignment
- Two vs. many classes
- Equal vs. unequal misclassification costs
- Assigning class priors (π_J) vs. using the proportions in the data

Global vs. local

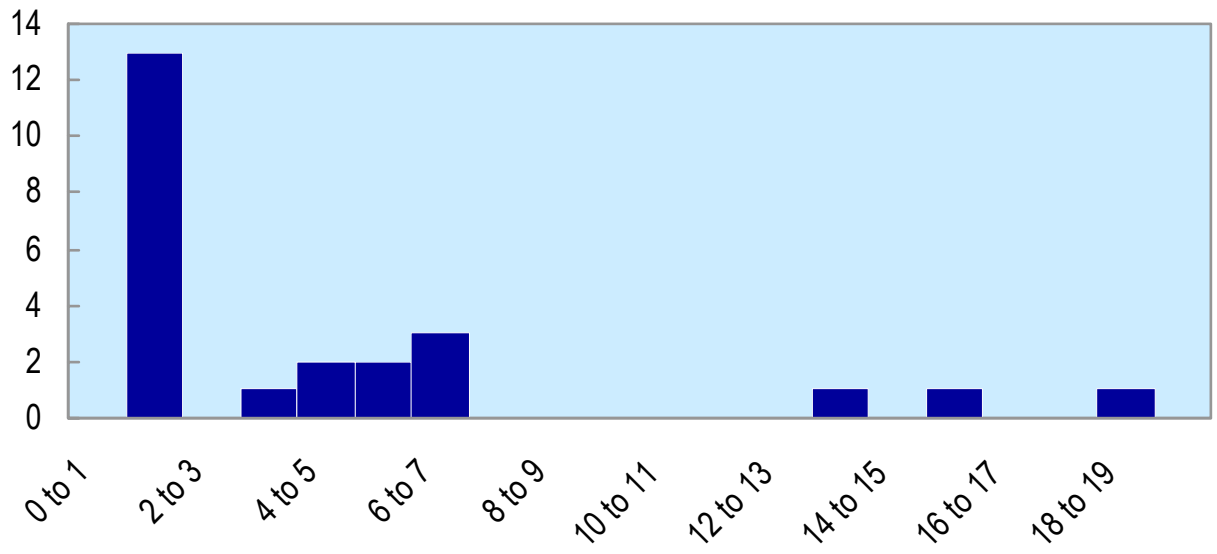
Most Global Model Imaginable	Most Local Model Imaginable
■ Unweighted one parameter model	■ Nearest neighbor
■ High bias, low variance	■ High variance, low bias
■ Appropriate if low SNR	■ Appropriate if high SNR

Density estimation in classification problems

- A statistical problem in its own right …
- … but also a way of handling the classification problem
- If you have populations A and B
  - Estimate the densities $f_A(x)$ and $f_B(x)$
  - Estimate the prior class probabilities π_A and π_B
  - Then assign a new observation with coordinates x to the class J that maximizes $\pi_J f_J(x)$
  - The prior probabilities can be taken from the data or from other knowledge
  - Estimating the densities is the tough part

Density estimation

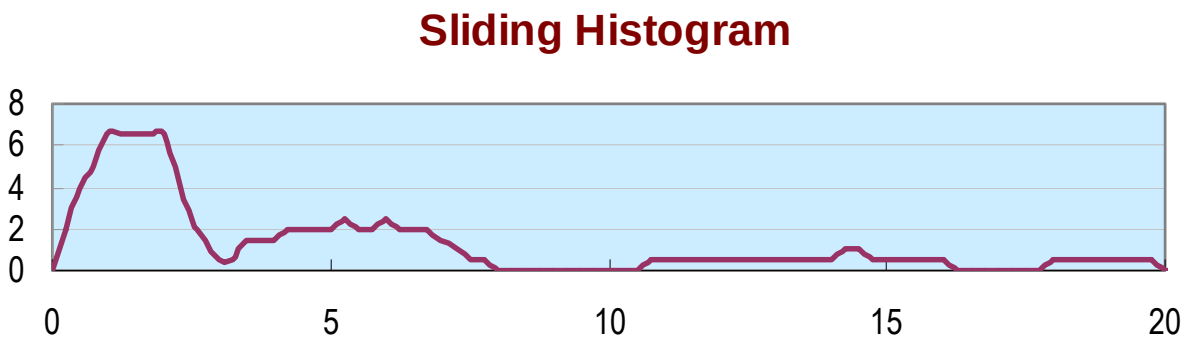
- Simplest density estimator is a histogram
- Can generalize this by a sliding histogram: Height at any one point depends on number of observations within a specified distance
- More generally, can use ‘kernel’ functions to take weighted averages, giving more weight to nearer points



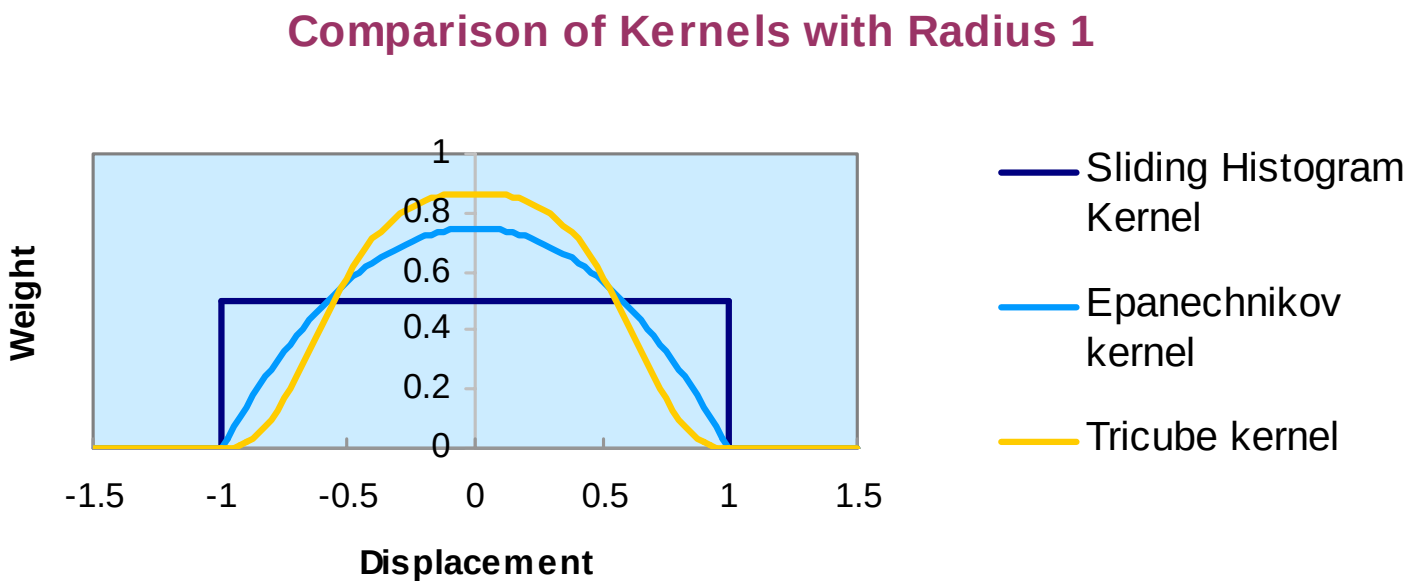
1.06	3.96
1.06	4.39
1.09	4.45
1.12	5.04
1.27	5.88
1.30	6.12
1.40	6.32
1.47	6.87
1.57	11.72
1.69	13.68
1.77	15.17
1.82	19.00
1.86	

Density estimation — kernels

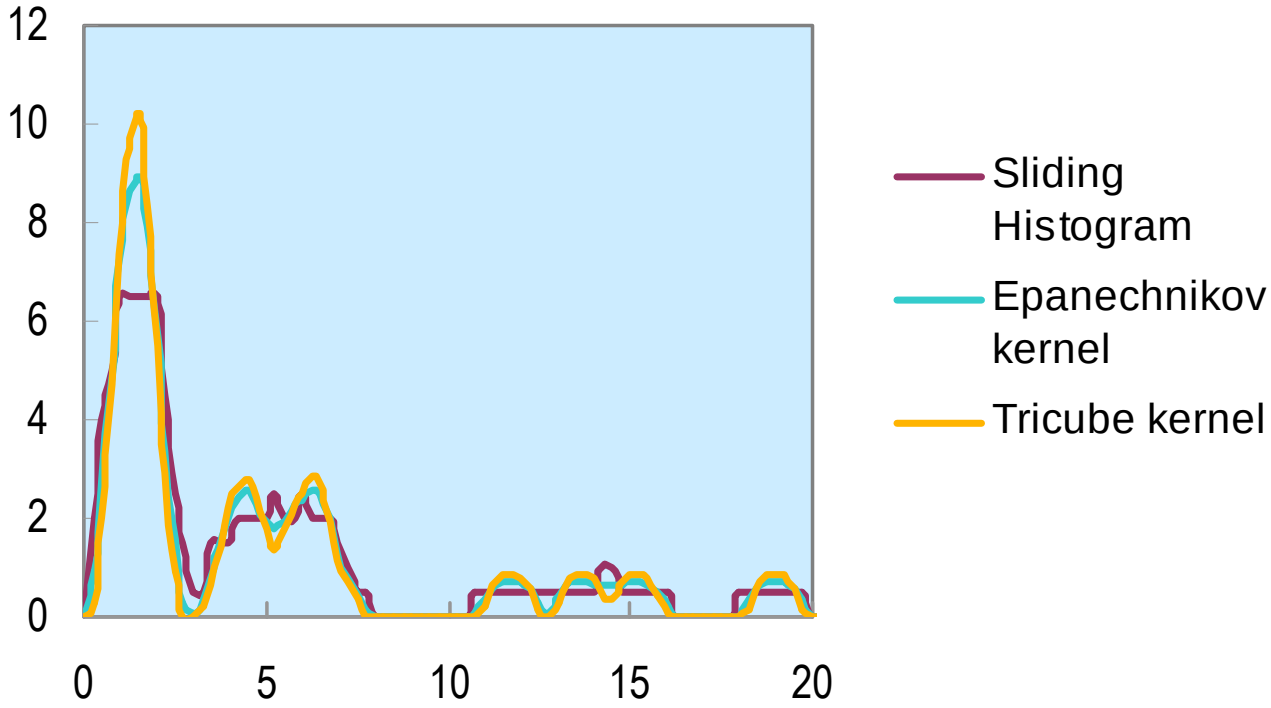
- "Sliding histogram" is a kernel where the kernel drops off from 1 to 0 at a specified distance
- Common choices of kernel (with kernel radius of r, and object at distance d)
  - Epanechnikov $1-(d/r)^2$ minimizes mean square error asymptotically
  - Tricube $(1-(d/r)^3)^3$
  - Can use a normal distribution, but it does have infinite radius … often undesirable
- Note that endpoints are a problem. Extrapolation is an extreme problem



Density estimation — kernels



Density estimation — kernels



Some techniques

- Naïve Bayes
- K nearest neighbors
- Discriminant Analysis (Linear, Quadratic …)
- Logistic Regression (various links …)
- Trees (e.g., CART, CHAID, C4.5)

Naïve Bayes

- There are a lot of refinements to naïve Bayes, but the basic idea is very simple, and is also known as “idiot’s Bayes”:
  - Assume there are no interactions
  - Model densities univariately
    - Use contingency table for discrete predictor
    - For continuous predictor, usually bin the variable to make it discrete, but could just as easily use a kernel density estimator
  - In form, looks like a generalized additive model, except the additive bits often much simpler than splines
    - But much faster and simpler to fit

Naïve Bayes

- Example
  - Suppose there is a population of 100 men and 50 women
  - Of the population, 20 of the women are wearing skirts. The other 130 are wearing pants
  - Of the population, 20 of the men and 30 of the women have hair shoulder-length or longer
  - The goal is to predict gender from the other observations
- Naïve Bayes assumes that for each class, the densities are the products of the marginals

Men	Total	Long Hair	Short Hair
Total		20%	80%
Pants	100%	20%	80%
Skirt	0%	0%	0%

Women	Total	Long Hair	Short Hair
Total		60%	40%
Pants	60%	36%	24%
Skirt	40%	24%	16%

Naïve Bayes

100 Men	Total	Long Hair	Short Hair
Total		20%	80%
Pants	100%	20%	80%
Skirt	0%	0%	0%

50 Women	Total	Long Hair	Short Hair
Total		60%	40%
Pants	60%	36%	24%
Skirt	40%	24%	16%

- Predicted probability of observed person being male, assuming class priors in the data

Prob(Male)		Long Hair	Short Hair
Pants		52.6%	87.0%
Skirt		0%	0%

- Predicted probability of observed person being male, assuming equal class priors

Prob(Male)		Long Hair	Short Hair
Pants		35.7%	76.9%
Skirt		0%	0%

Naïve Bayes

- What naïve Bayes does not take into account is that the second square of data could actually look like
- This would change the resulting probabilities considerably
- For example, with equal priors, the probability of someone being male given long hair and pants would be 50%, but naïve Bayes would still predict 35.7%

50 Women	Total	Long Hair	Short Hair
Total		60%	40%
Pants	60%	20%	40%
Skirt	40%	40%	0%

Naïve Bayes

Advantages

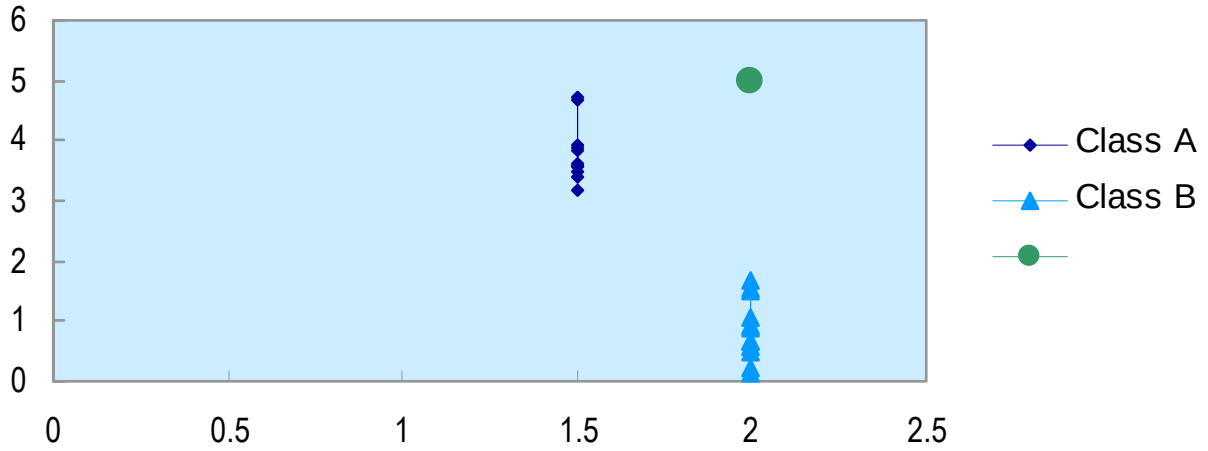
- Easy to do
- Very easy to interpret
  - Just one-way tables put together
- Decision boundaries fairly flexible but not completely general

Disadvantages

- Sensitive to feature selection
  - Easy to double count effects
  - On the other hand, can automate feature selection and make naïve Bayes a good method even on problems with more predictors than observations
- Does not handle interactions gracefully

Discriminant analysis

- How to group things?
- Naïve approach:
  - For each class, take the centroid of the training data for that class
  - Classify new points to the closest centroid
- What's wrong with this?
  - Define "close"
  - Normalizing predictor variables won't help (much)
    - Differences in some may be more important than differences in others
    - Some may be strongly correlated

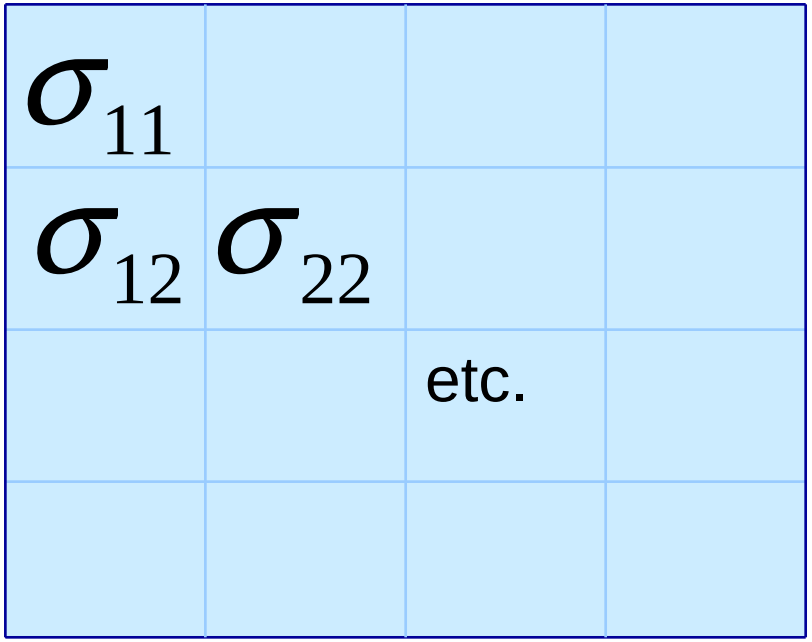


Linear Discriminant Analysis (LDA)

- Normal distance works well for spherical clusters
- To the extent that classes are not spherical, rescale them
- Modeling each class with a multivariate normal does three things:
  - Centers class density at centroid
  - Accounts for elliptical distribution
  - Accounts for dispersion of each class
- But … tons of parameters to estimate:
  - If p predictors and k classes, then
  - k p-dimensional centroids and k pxp covariance matrices
- Simplify:
  - Assume each class has same covariance matrix

Linear Discriminant Analysis (LDA)

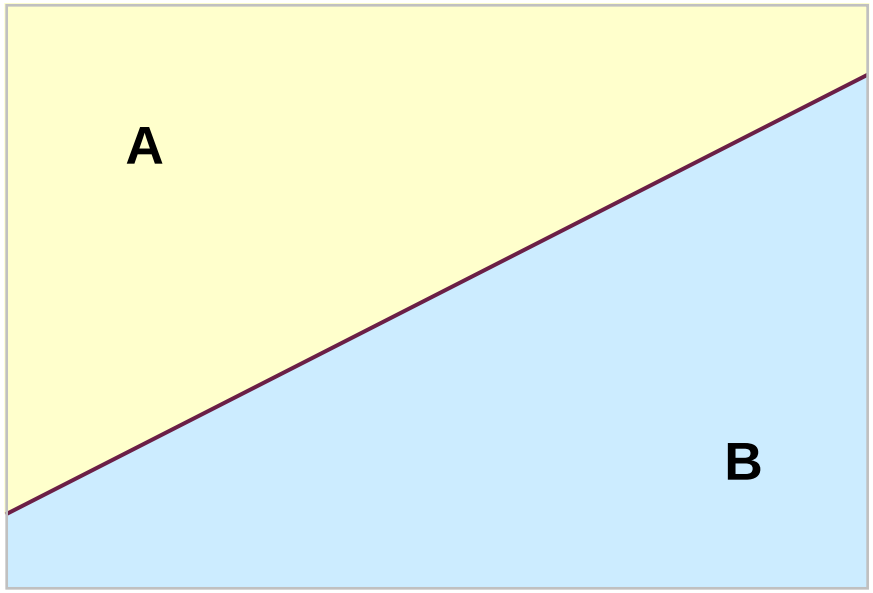
- Estimation
  - Estimate centroids
  - For each observation (x,J), with class centroid C_J, consider x-C_J
  - Determine the covariance matrix of the x-C_J
    - Easy enough to do one pair of coordinates at a time: Covariance is just the average of the product less the product of the averages



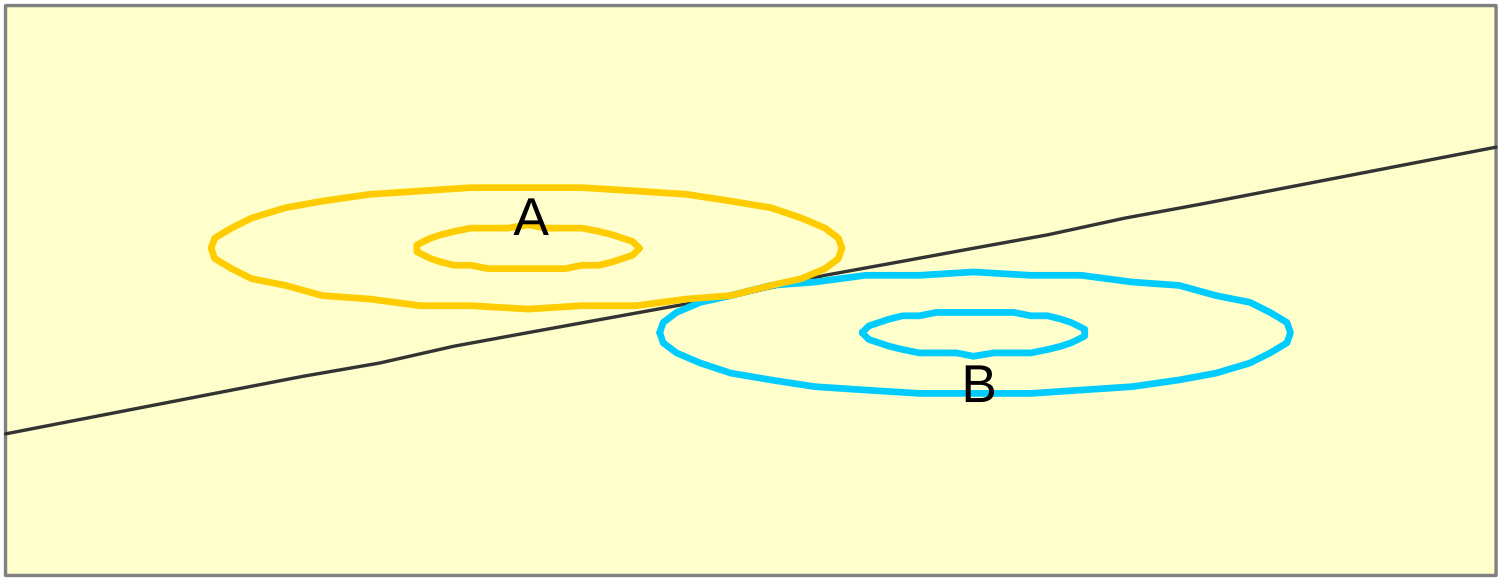
- The result is called linear discriminant analysis because the decision boundary will be linear (in fact, a hyperplane)
- Why?
- Because a linear transformation will make the ellipsoids into spheres (when we know the boundary is a hyperplane)

Linear Discriminant Analysis (LDA) virtues

- There are really fewer degrees of freedom than it appears
  - Decision surface is a hyperplane in predictor space, so only p+1 degrees of freedom for 2 class problem if p is the number of predictors
  - The decision surface for a 2 class problem is the same as that resulting from linear regression
    - Thus, it is not silly to apply linear regression to 2 class problems

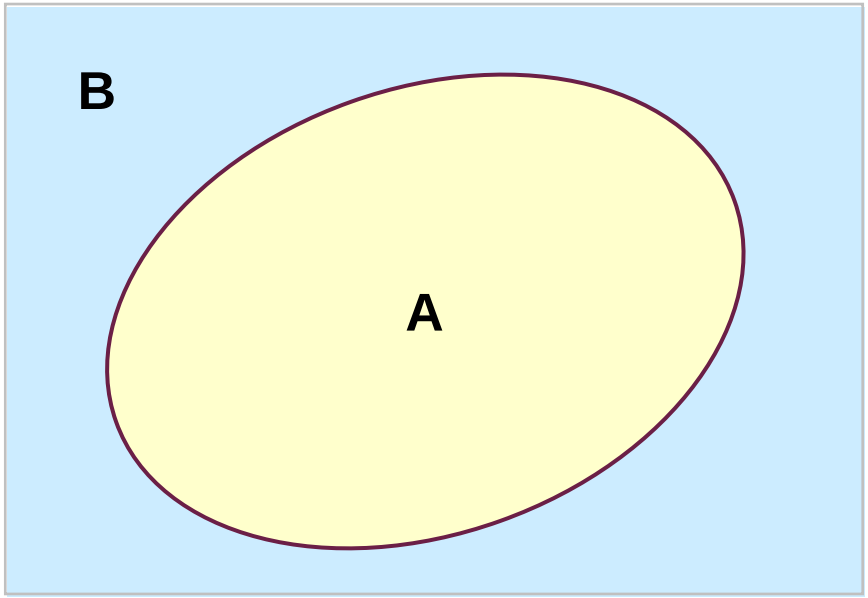


LDA with level curves of densities



Quadratic Discriminant Analysis (QDA)

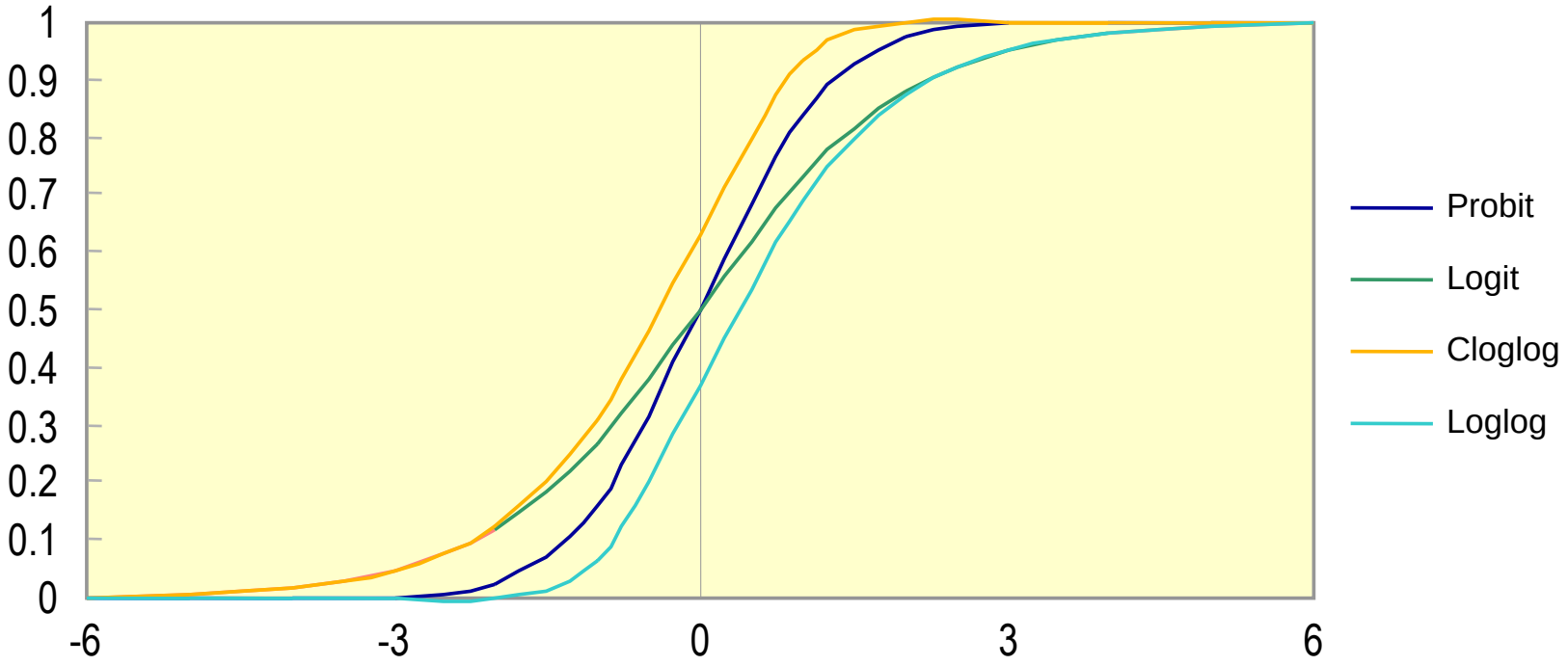
- If you have a ton of data, you can try estimating each covariance matrix separately
- Not only a lot of parameters to estimate
  - … but also more sensitive than LDA to non-normality
- Harder to interpret … decision surface is not linear
- Poor method if any class has few representatives, no matter how huge the data set



Logistic regression

- Generalized Linear Model
- Dependent variable is conditionally Bernoulli (0 or 1)
  - Note that you cannot think of this as “Bernoulli errors”
- Various ways of handling more than 2 classes
- The modeled probability of success given $x_1, \ldots, x_n$ is
  - $f(b_0+b_1x_1+b_2x_2+ \ldots +b_nx_n)$
- Note that this always gives a decision boundary linear in the x_i
- Choices for f(z):
  - Cumulative normal Φ (also called “probit” regression)
  - Logistic function: $e^z/(1+e^z)$
  - Complementary log-log: 1-exp(-exp(z))
  - Log-log: exp(-exp(-z))

Logistic regression



Logistic regression

- Logistic link
  - Use this one unless you have a reason to do otherwise
  - Can interpret $b_0+b_1x_1+b_2x_2+ \ldots +b_nx_n$ as the log of the odds of success
  - Note: probability of success = odds of success / (1 + odds of success)
  - Effectively a multiplicative model for the odds
    - Interpret the b_i
    - Allows for “retrospective” or stratified sampling
      - Because sampling does not change the relative odds
      - So it won’t bias the answer … it just changes the intercept
- Logistic does not like to predict pure answers (predictions near 0 or 1)
  - Probit loves to do this
  - Logistic preferable if there’s “always a chance” anything might happen
- Complementary log-log looks very similar to logistic for rare classes
  - Not appropriate if successes are common
- Log-log not appropriate if failures are common

Logistic regression

Advantages

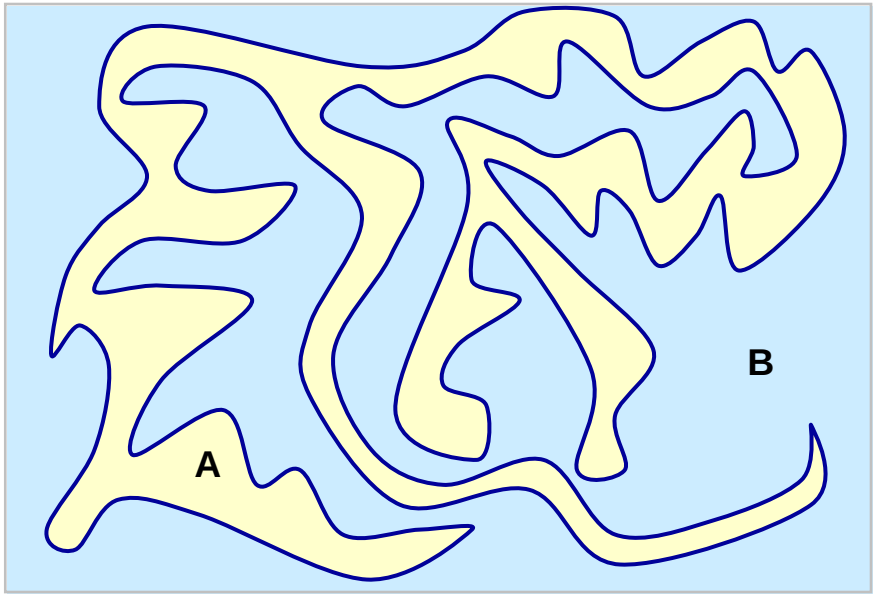
- Scores interpretable in terms of log odds
- Constructed probabilities have chance of making sense
  - Modeled directly rather than as ratio of two densities
- A good “default” tool to use when appropriate, especially combined with feature creation and selection

Disadvantages

- Invites over-interpretation of parameters
- For example, if a 10% rate increase
  - Causes lapse rates for customers under age 30 to increase from 15% to 20%
  - Causes lapse rates for customers 30 and over to increase from 5% to 10%,
- Then logistic regression says the older customers are more price sensitive
  - Their odds of lapse increased by a factor of 19/9
  - The young customers odds of lapse increased by a factor of 17/12
- Doesn’t generalize to 3+ classes as painlessly as LDA

k nearest neighbors

- Score each observation by vote of the nearest k training points
- Traditional for k to be odd
- Note that is k=1 then the training error will be 0 by definition
  - This is **not** necessarily a good thing
  - Cross-validation will give good estimate of what the error would be on a test set, assuming independent observations in the training set
- This is very similar to kernel density estimation
  - But the neighborhood size is determined by the density of observations
  - Within the neighborhood all observations count equally



k nearest neighbors

Advantages

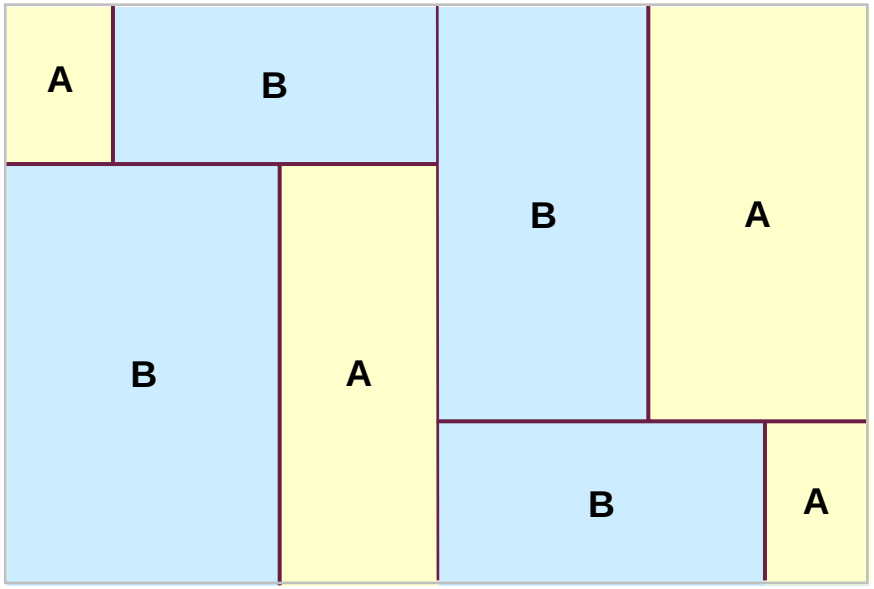
- Very flexible … can model almost any decision boundary
- Requires no distributional assumptions

Disadvantages

- Computationally painful
  - Search entire training set for nearest neighbors for **every** test point
  - There are ways to speed this up, but still slow
- Breaks down with large number of predictors (curse of dimensionality)
- Too flexible
  - Easy to overfit
  - Of course, can usually cure this by choosing k large enough
  - Often k=1 is terrible
- Need to decide how to scale the axes
  - Standardizing variables is not necessarily a sensible solution

Decision trees

- Recursively split the data
- Greedy
  - At each iteration choose split to maximize some measure of significance or purity
  - Continue until reaching some stopping criterion, e.g.,
    - Don't split nodes smaller than a certain size
    - Don't split nodes with significance less than a certain amount
  - Prune this back
- For a continuous model, each box would be labeled with a score rather than a letter



Decision trees (CHAID)

- Some common algorithms
  - CHAID
  - CART
  - C4.5
- CHAID (d categories of dependent variable)
  - Classify predictors as ordinal or categorical
  - For each categorical (resp., ordinal) predictor, merge the pair (resp., adjacent pair) of categories where the 2 x d contingency table is least significant, if it is not significant at a certain level p
    - A missing value can be considered adjacent to any value
    - Alternate this with testing whether merged categories can be split at that significance level. If d=2, can treat categorical predictors as ordinal, ordered by the proportion of the first class
    - Sum of [(observed – expected)^2/expected] is chi-square with (d-1) degrees of freedom
  - This is just like stepwise regression (using chi-square instead of F tests)

Decision trees (CHAID)

- Eventually, one has determined how to merge the categories for that predictor
  - If there are c of them, now compute the significance level of the c x d contingency table, which is chi-square with (c-1)(d-1) degrees of freedom
  - Bonferroni adjustment: **multiply** this significance level by a penalty for the number of ways the original classes could have been collapsed into c classes
- Repeat this process for all predictors
- Split on the most significant predictor
- CHAID as such has a stopping rule but no pruning rule
  - However, could always allow a generous significance level (to overfit) and then prune as per CART

Regression trees (CHAID)

- For a continuous dependent variable, replace the chi-square test with an F-test for including dummy variables in a linear regression
- Note that this is very similar to selecting the **first** variable to include in a stepwise regression
- Additional consideration of Bonferroni adjustment and determining the number of ways the split should divide
- Main difference from stepwise is one then divides the dataset

Decision trees (CART)

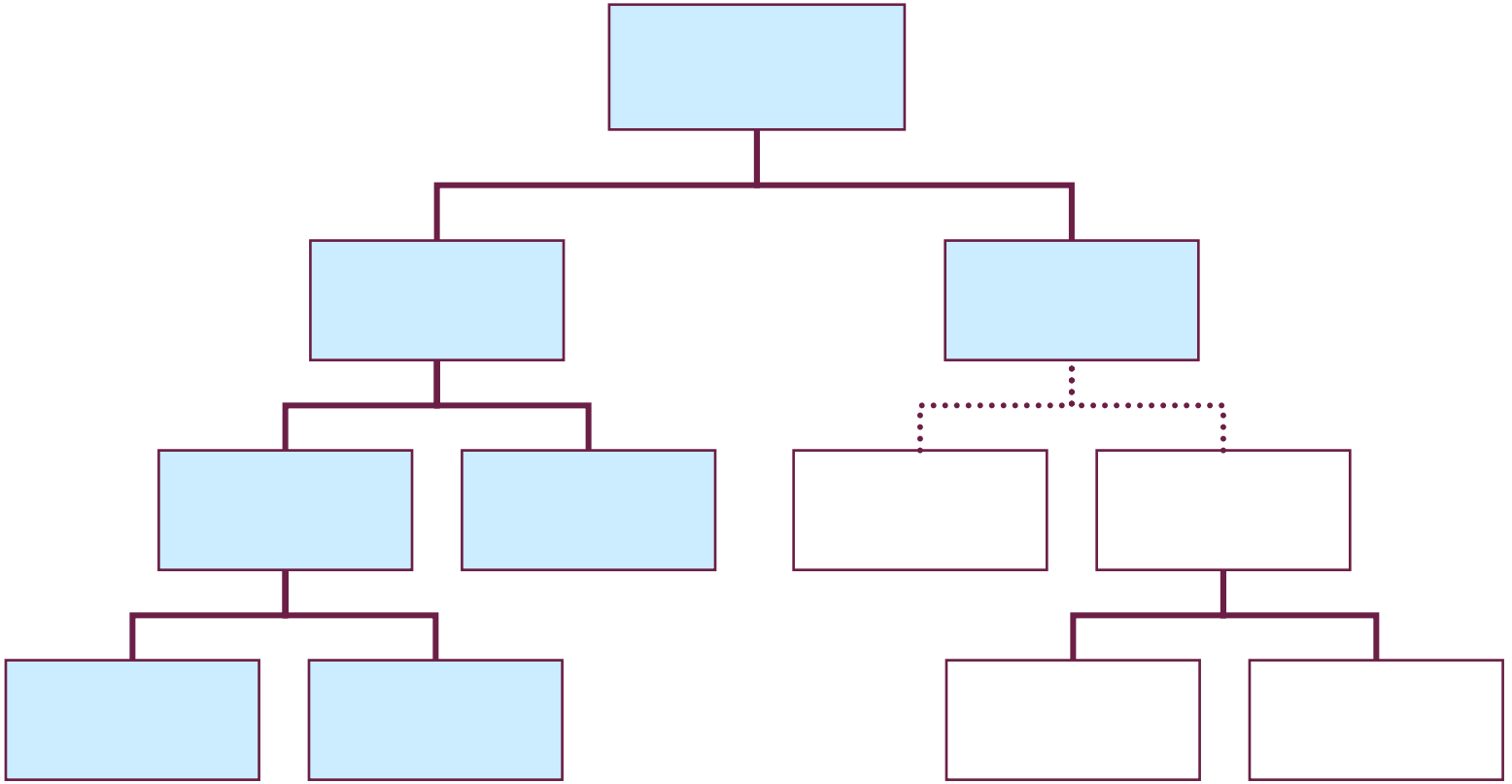
- Consider all **binary** splits on all predictors (splits of the form x>a for ordinal variables)
- Various different criteria for determining the best split, will focus on Gini criterion:
  - Minimize expected misclassification cost
  - Sum of misclassification costs for each child node
    - If the left child node has probabilities 90% A, 5% B, and 5% C, and takes 30% of the observations
    - And the right node has probabilities 20% A, 70% B, and 10% C, and takes 70% of the observations
    - And the cost of misclassifying an A or C object is 1, but the cost of classifying B as A is 2 and B as C is 3, then
    - The total misclassification cost for the left node is:
      - 30% of 90%*5%+90%*5%+2*90%*5%+3*5%*5%+90%*5%+90%*5%
    - Compute for right node similarly and add

Decision trees (CART)

- Grow an extremely overfit model (large tree)
- Determine an order in which to prune back
  - Score each prune as
    - (increase in in-sample misclassification cost) / (decrease in number of terminal nodes)
    - Note that this can be seen as requiring a minimum usefulness for each degree of freedom
- Obtain a series of pruned trees, each corresponding to a required "usefulness" per node (marginal utility per node of every "prunable subtree" [for example the pink nodes on p. 70] must meet the standard)
- Use cross-validation to determine which in the series of pruned trees is the best
  - By determining the optimal value of the tuning parameter (required usefulness per node)

Regression trees (CART)

- Splitting rule is the binary split that minimizes within class variance
- In the pruning step, increase in in-sample squared error replaces increase in in-sample misclassification rate



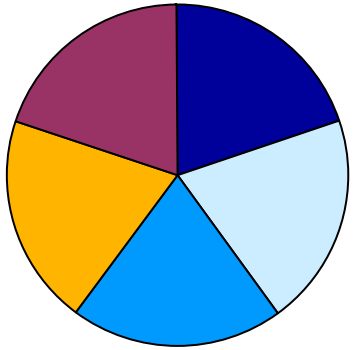
Model Validation

Training data, test data, and hold-out data

- If you fit a model to all the data, left with no way to validate it
- A model fit to part of the data can be tested on the rest of the data
- If you use test data to compare models that have been built on the training data, then you have to some extent fit the test data also
- So need a third set, hold-out data, with which to derive an unbiased measure of how good the model is
- Shortcuts:
  - Compare models on training data only
    - — Then can use test data as hold-out data also
  - Use cross-validation if must model all the data [due to]
    - — Regulation
    - — Low SNR
- Validating on out-of-sample data gives freedom in developing technique
  - Don’t need to worry whether nice statistical tests (which require lots of often unrealistic assumptions anyway) are available for in-sample data
  - Just validate!

Cross-validation

- For estimating accuracy
  - Divide the data into N (N=5 commonly) equal parts
  - On each 4/5 of the data, fit a model by exactly the same process used for the full data set
  - Use, for example, the model that has not “seen” slice 1 to score slice 1. Score the entire dataset in the way and measure the error
  - This is an estimate of the prediction error
- Can also be used to tune a parameter or a feature selection
  - [But then not also for accuracy estimation]



Scoring

- Often a two class model produces scores
  - Observations with scores greater than a certain amount are classified to A; the rest to B
  - The cutoff score can be changed
    - E.g., could use the cutoff that gives the lowest misclassification cost
- A soft assignment model is a model where these scores can reasonably be interpreted as probabilities

Some terminology

Confusion Matrix

Actual Class	Predicted Class		
	A	B	C
A	122,332	34	322
B	3,124	214,324	2,345
C	312	345	23,445

Some terminology (ROC curve)

- Receiver-Operating Characteristic curve (ROC curve):
  - True positive rate vs. false positive rate
  - Allows comparison of several types of model each tuned to various false positive rates by changing the misclassification costs
- Area under the ROC curve is a commonly used comparison (more is better)
  - As with all tests, comparison should be on test data, not training data

Some terminology

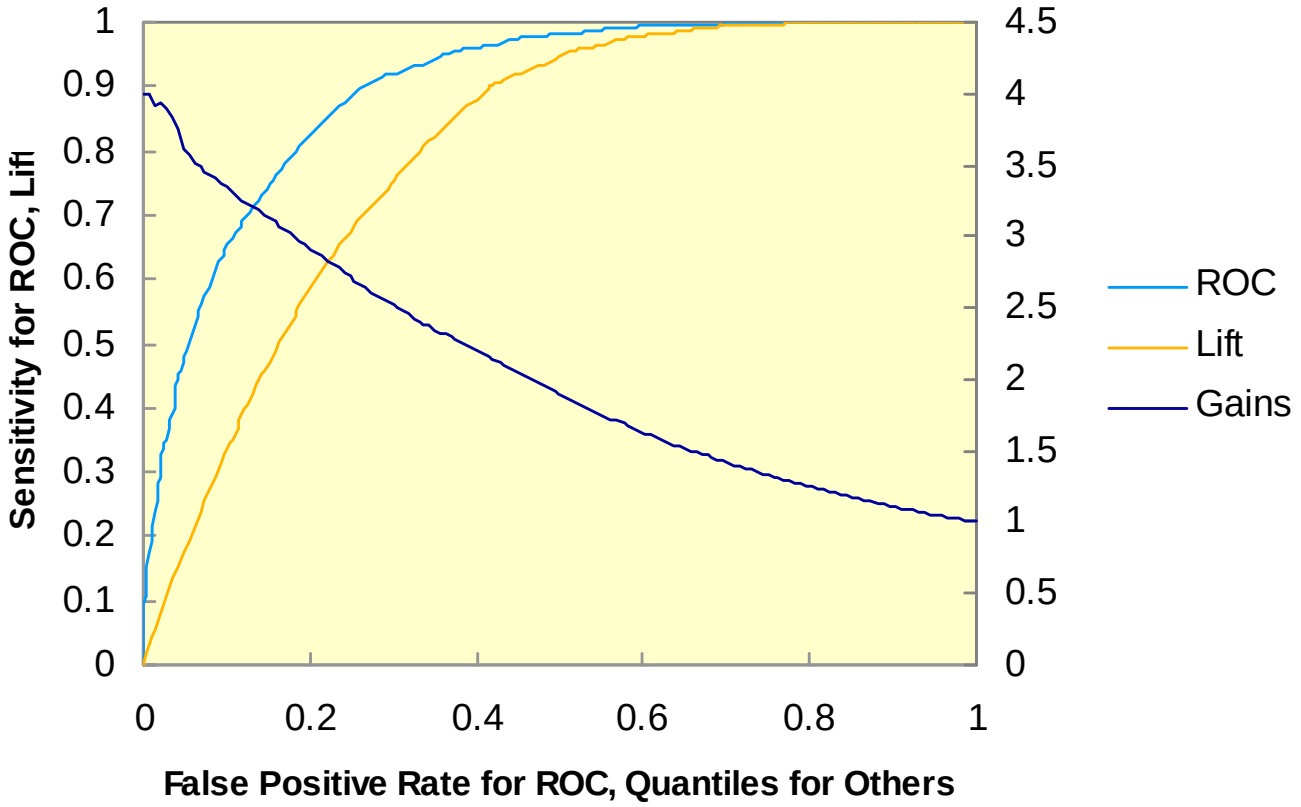
Gains Chart

- With scores, can vary number classed as type A continuously. Call this x%
- Gain = proportion of those classed as A that are A compared to proportion in general population that are A
- Gain is >= 1 and is decreasing as one moves to the right (including more quantiles in the mailing, for example). Flat line at 1 is worthless model
- Often used in response modeling: The “gain” vs. random mailing

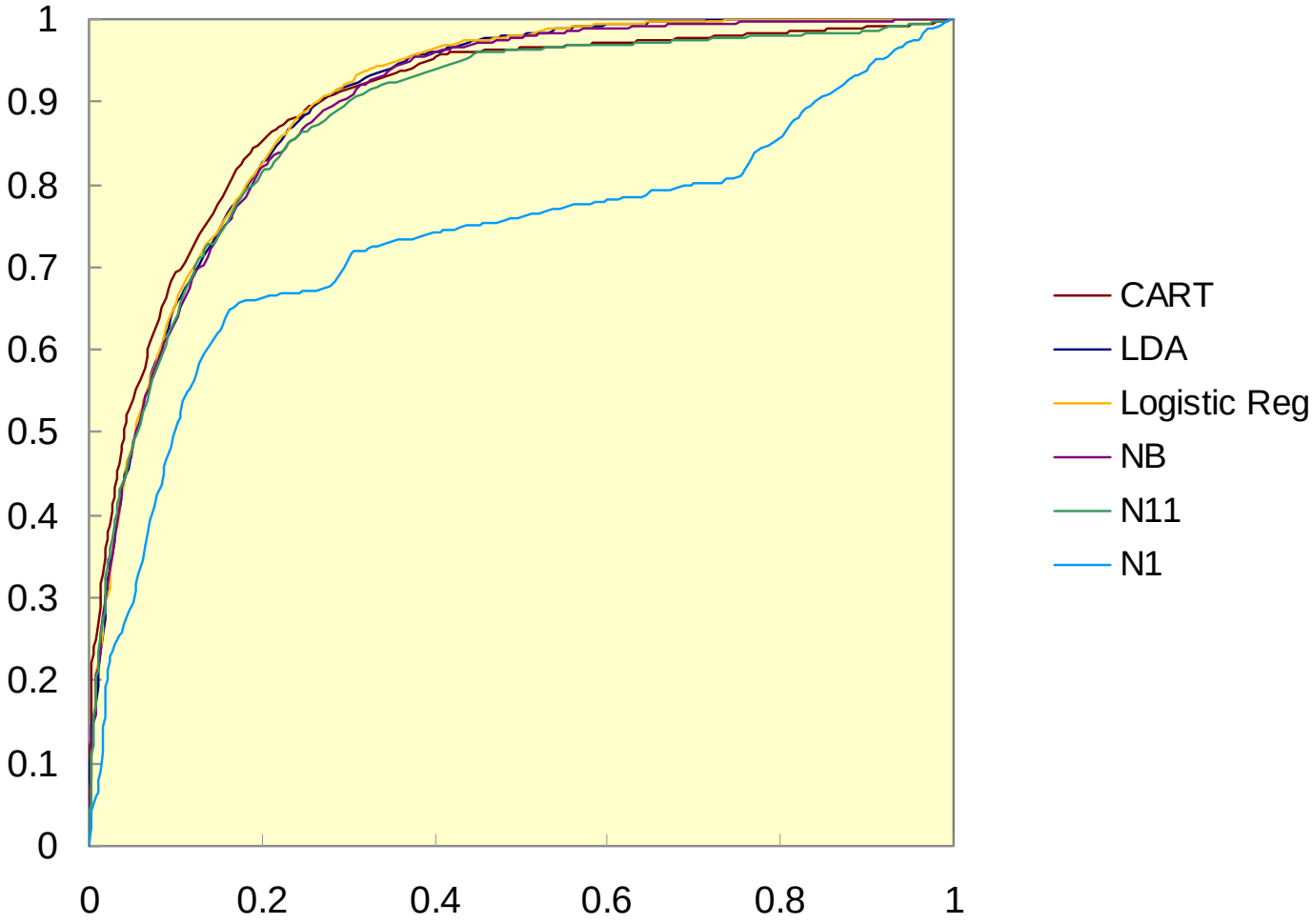
Lift Curve

- Lift = percent of class A that falls in the first x% of scores
  - True positive rate as a function of quantile
  - If you have the 20% scored most likely to be in class A, then false negative rate will be less than 20%, so lift curve is to the right of ROC curve

Some terminology



ROC curve comparison



Some terminology — continuous models

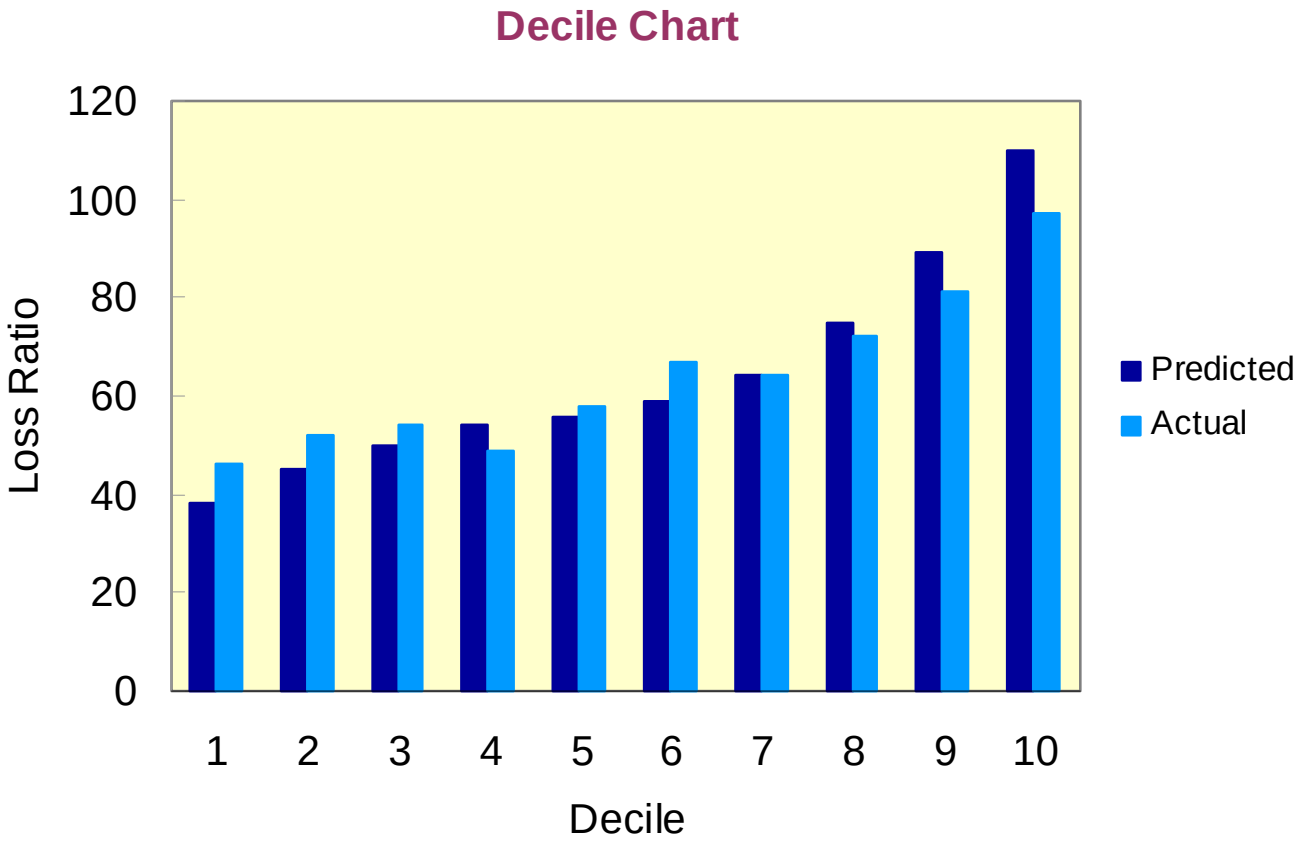
Decile Chart

- Sort the out-of-sample data by the predicted value
- Show the actual dependent variable mean for each decile (or twentieth, or whatever the appropriate bucket size is)
- Similar to a gains chart except that one shows the mean for each decile, rather than for the first, the first two, the first three, etc.

Lift Curve

- Can do these for continuous data also
- E.g., the wealthiest 5% have 90% of the money
  - Except, in this case, it's the 5% your model thinks are the wealthiest
  - If it's a bad model, they might only have 6% of the money

Some terminology



Regularization

Building a model — regularization

- Idea: Compromise between a simple and a complex model by introducing a tuning parameter to average them in some way
- Use cross-validation to determine the appropriate value of the tuning parameter
- Examples:
  - For discriminant analysis, can directly average QDA and LDA covariance matrices
  - For linear regression (or GLM), can penalize large parameter values
  - Credibility …can use cross-validation to optimize K
  - A CART prune
    - The tuning parameter is the cost of adding a node

Building a model — regularization

- How to get the right amount of flexibility
- Average local covariance estimate with global one (typical actuarial thing to do)
- Two types of averages suggested by Friedman:
  - Average class covariance matrices with grand mean
    - $\Sigma_{J,Z} = \Sigma_J Z + \Sigma (1-Z)$
    - Choose Z by whatever produces the best fit
    - Ideally in terms of cross-validation
  - Average the resulting covariance matrices with scalar multiple of identity
    - Choose the scalar multiple to have the same trace as $\Sigma_{J,Z}$
    - Scaling of predictors suddenly matters
    - Be careful with this if you have collinearity, since this assumes there isn't much collinearity

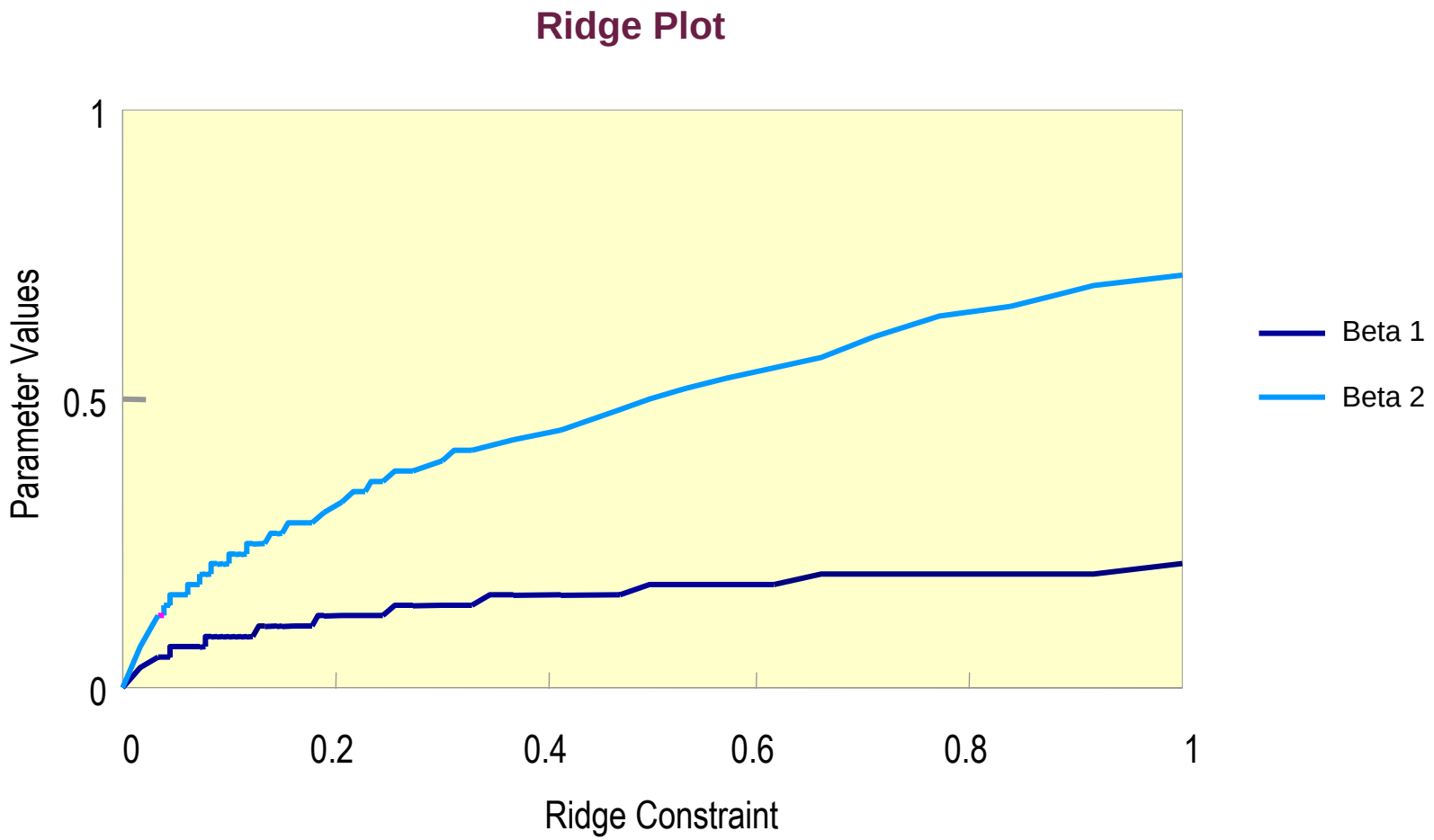
Building a model — regularization

SHRINKAGE

Ridge Regression

- Used to penalize large parameters, using sum of squares of parameter sizes as the penalty
- Center Y
- Center **and standardize each** X_i (divide by standard deviations), separately for each i
- Equation to minimize is
  - $\Sigma_i(y_i-\Sigma_j\beta_j x_{ij})^2$ subject to $\Sigma_j\beta_j^2 < \Lambda$
    - Equivalent to minimizing $\Sigma_i(y_i-\Sigma_j\beta_j x_{ij})^2 + \lambda\Sigma_j\beta_j^2$ with $\lambda > 0$
- One use of ridge regression is to control for multicollinearity
  - This is the reason for standardizing the predictor variables

Building a model — regularization

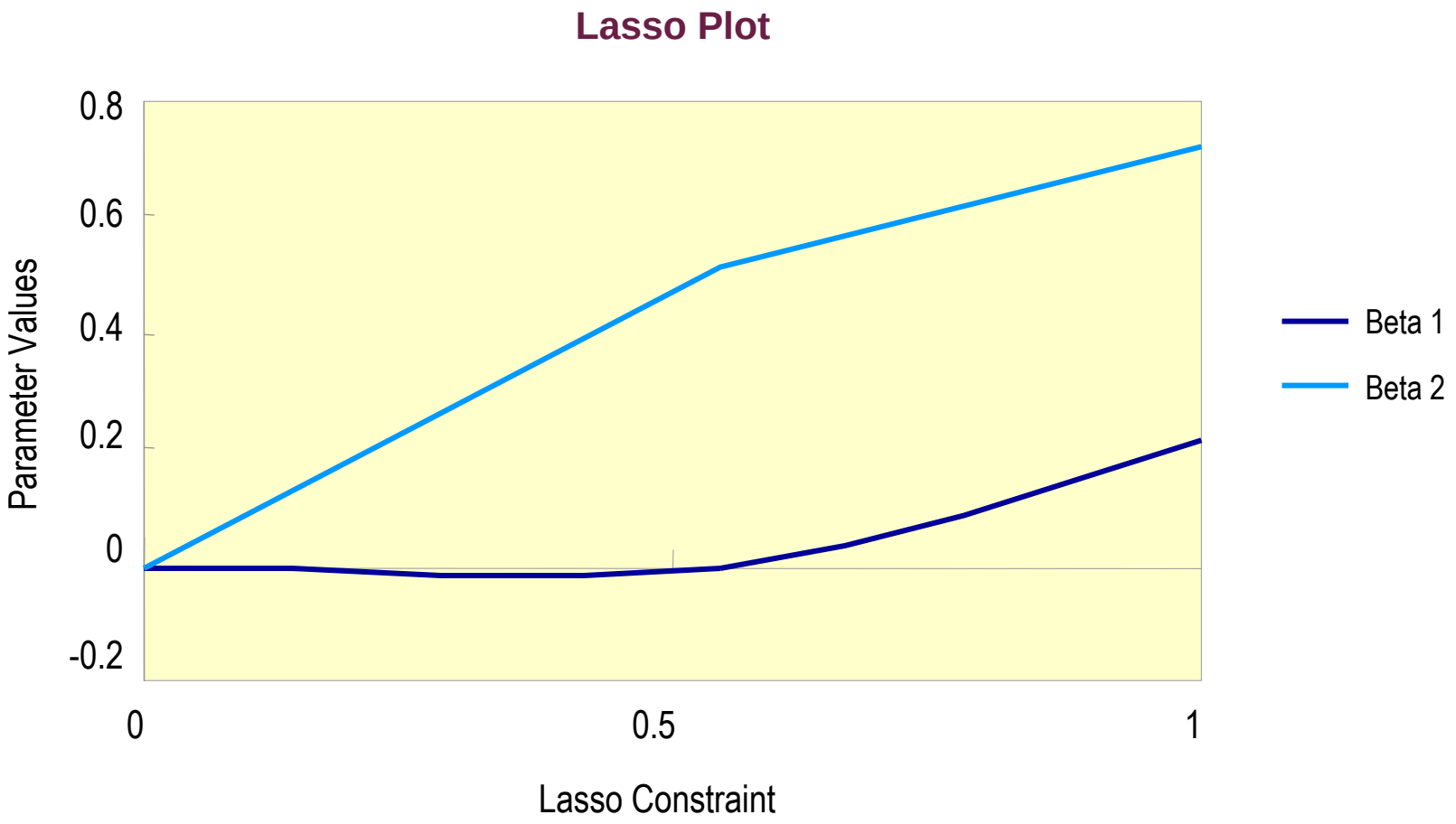


Building a model — regularization

The Lasso

- Unlike ridge, can shrink some parameters all the way to 0
- Penalty is sum of absolute parameter values, i.e., minimize
  - $\Sigma_i(y_i-\Sigma_j\beta_j x_{ij})^2$ subject to a constraint $\Sigma_j|\beta_j|<\Lambda$
  - This corresponds to minimizing $\Sigma_i(y_i-\Sigma_j\beta_j x_{ij})^2 + \lambda\Sigma_j|\beta_j|$ with $\lambda>0$
- In Bayesian interpretation, corresponds to prior for each β_i that is double exponential with a density of $(\sigma^2/\lambda)\exp(-\lambda|\beta_i|/2\sigma^2)$
  - Note that $\text{var}(\beta_i)=4\sigma^2/\lambda$. Call this τ^2
  - This corresponds to a more diffuse (more tail-heavy) prior than the normal
- Again, in "standard" penalized regression, one centers Y and centers and standardizes each X_i.

Building a model — regularization



Useful references

Brieman, Friedman, Olshen, and Stone, *Classification and Regression Trees,* Chapman & Hall, 1984

Domingos, Pedro, “The Role of Occam’s Racor in Knowledge Discovery”, *Data Mining and Knowledge Discovery*, 3, 409-425, 1999

Hand, David J., *Construction and Assessment of Classification Rules,* Wiley, 1997

Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining Inference and Prediction,* Springer, 2001

Hastie and Tibshirani, *Generalized Additive Models,* Chapman & Hall, 1990

McCullagh and Nelder, *Generalized Linear Models*, 2nd ed, Chapman & Hall, 1989