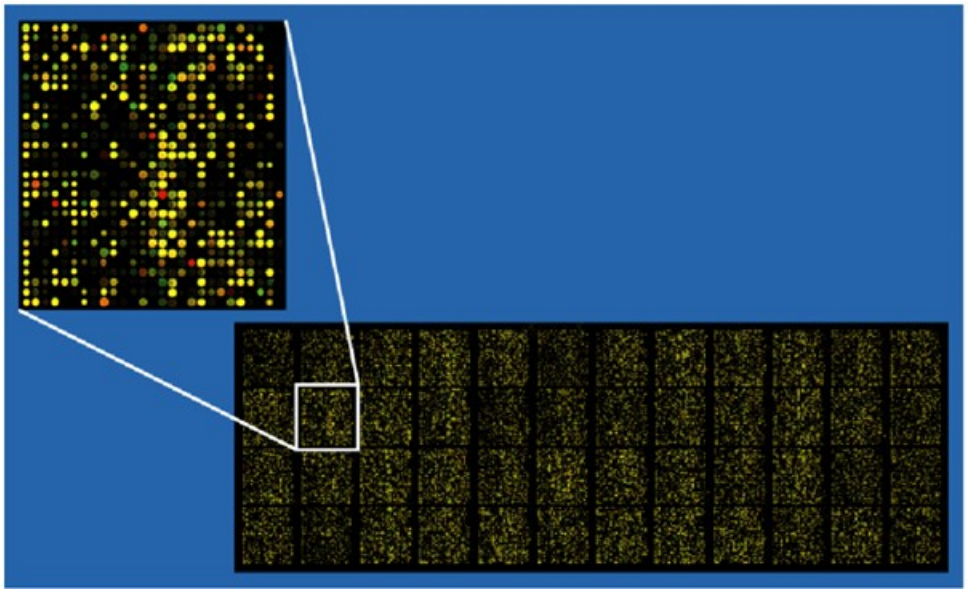
On the Chance Accuracies of Large Collections of Classifiers

Mark Palatucci and Andy Carlson
Carnegie Mellon University
July 7, 2008

High Dimensional Classification Problem

- Example: Gene studies using microarrays

Given a microarray of a gene expression levels, which patients are more likely to develop cancer?

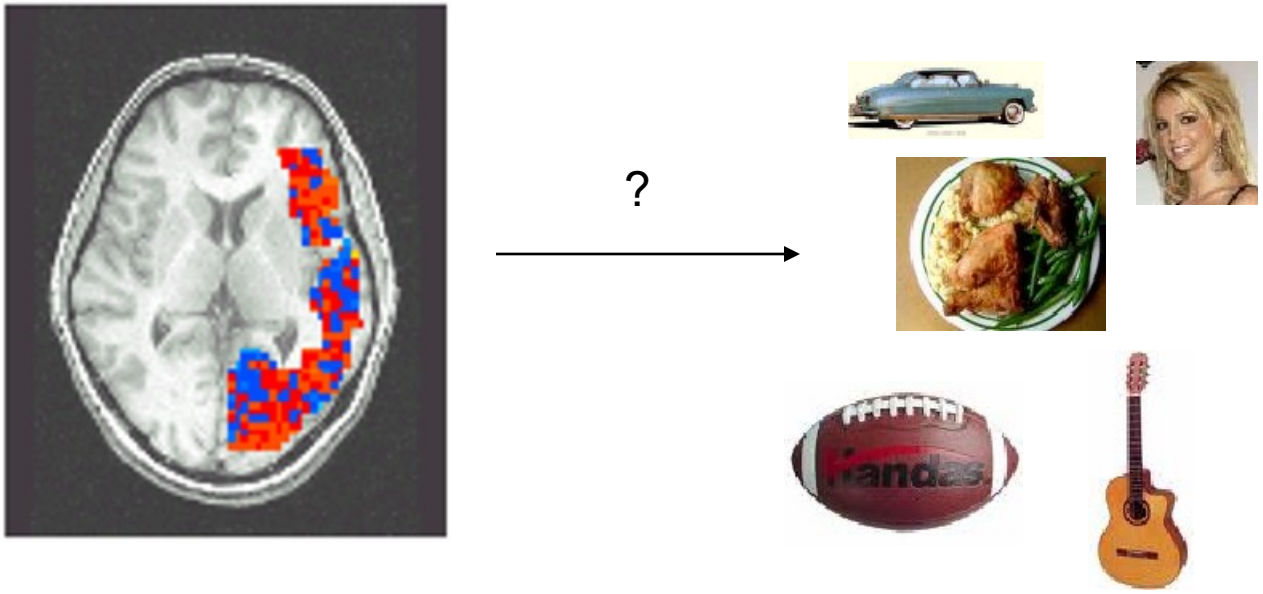


Gene classification tasks can have thousands of raw features and only ~100 training examples.

High Dimensional Classification Problem

- Example: Cognitive state classification

Given a functional magnetic resonance image (fMRI) of a person's neural activation, what is the person thinking about?

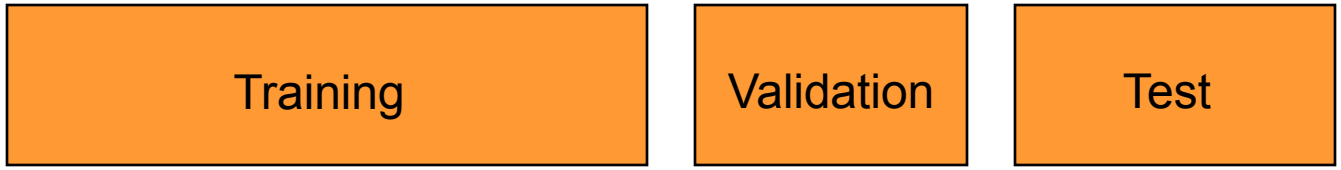


fMRI classification tasks can have hundreds of thousands of raw features and only ~100 training examples.

Feature Selection in fMRI

- Common fMRI feature selection methods
  - Embedded (learner dependent)
    - L1 Regularized Logistic Regression
    - Support Vector Decomposition Machine (SVDM)
  - Filter (learner independent)
    - Average over spatial/temporal dimensions
    - PCA/ICA/Manifolds (for larger fMRI datasets)
    - Multiple hypothesis testing for active voxels (t-test, rank)
  - Wrapper (wraps around an induction algorithm)
    - Stepwise Selection (Forward, Backward)
    - Highest discriminating voxels on validation set

Discriminative Feature Selection



- Train a classifier using each feature individually
- Evaluate each of these single-feature classifiers on set of validation examples.
- Choose best performing features according to performance on validation set:
  - Choose top N
  - Choose all features that perform better than XX% accuracy

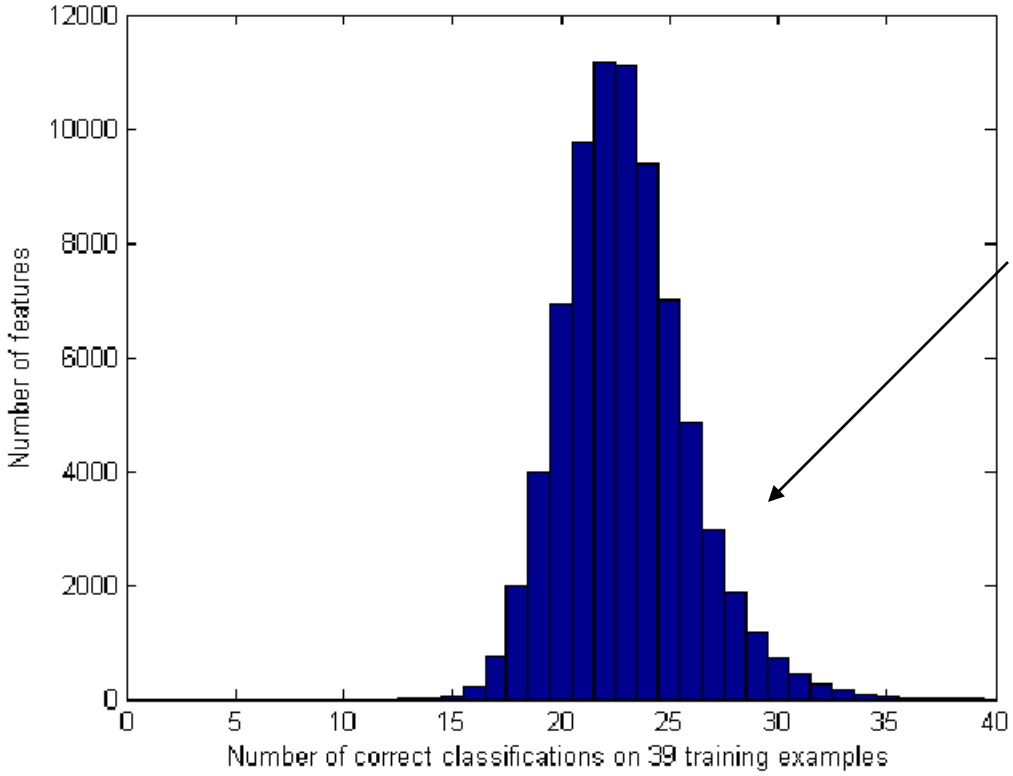
How do we choose a good ‘N’ or accuracy threshold?

Discriminative Feature Selection

- How do we choose a good ‘N’ or accuracy threshold?
  - Pick ‘N’ or accuracy threshold arbitrarily. E.g. N = 500 or accuracy >= 70%
    - Very dangerous with many features and small number of examples
  - Use a statistical hypothesis test
    - How do you choose alpha-level? E.g. 0.05?
    - How do correct for multiple tests? E.g. Bonferroni correction or false discovery rate (FDR)
  - Use a 2nd validation set to select
    - Computationally expensive
    - Difficult with so little data (e.g. less than 100 training examples)

Discriminative Feature Selection

- Simple experiment:
  - fMRI classification task: 80,000 features, 2 classes
  - Train a Gaussian Naïve Bayes classifier on each feature and evaluate on ~40 examples



What would be a good cutoff to use?

What would this look like if all classifiers were guessing randomly?

Feature Selection: Highest Chance Accuracy

Given M classifiers that each produce labels randomly for N examples, what is the expected accuracy of the *best* one?

Feature Selection: Highest Chance Accuracy

- Let X_i be the number of errors made by the i^{th} classifier:

$$X_1, X_2, \ldots, X_M \sim Binomial(N, p_{err})$$

- Order the samples:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(M)}$$

- The R^{th} smallest value is defined as the *R^{th} order statistic*:

$$X_{(r)}$$

Each order statistic is a random variable with its own distribution function

Feature Selection: Highest Chance Accuracy

- Let X_i be the number of errors made by the i^{th} classifier:

$$X_1, X_2, \ldots, X_M \sim Binomial(N, p_{err})$$

- The classifier with the smallest number of errors:

$$X_{(1)} = min(X_1, X_2, \ldots, X_M)$$

- Expected *minimum* number of errors:

$$\mathbb{E}\left[X_{(1)}\right]$$

Computing order statistic distributions and moments of continuous variables is easy. Discrete variables are much trickier.

Feature Selection: Highest Chance Accuracy

Theorem 4.1. *Highest Chance Accuracy*

$$\mathbb{E}[\mathcal{A}_H] = 1 - \frac{1}{N} \underbrace{\sum_{i=0}^{N-1} I_{p_{err}}(i+1, N-i)^M}_{\mathbb{E}[X_{(1)}]}$$

where $I_p(a,b)$ *is the incomplete beta function :*

$$I_p(a,b) = \frac{1}{\beta(a,b)} \int_0^p t^{a-1}(1-t)^{b-1} dt$$

N: Number of examples
M: Number of classifiers
p_{err}: Probability of a classifier making an error on an example

- Theorem defines a *natural significance* threshold

Feature Selection: Example

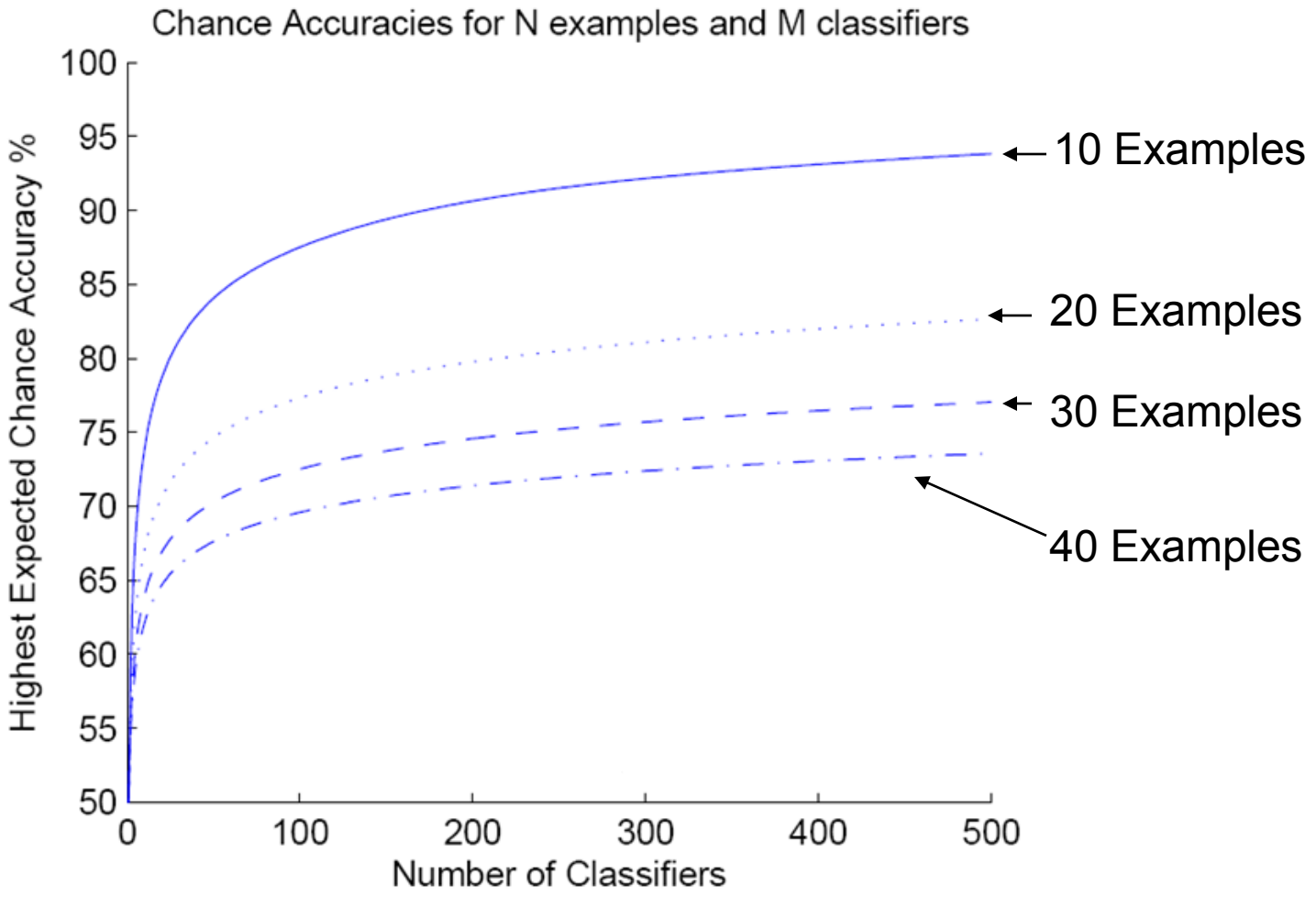
- Example 1: Predicting football games
  - Consider an office football pool with 200 participants betting (at random) on the outcome of 20 games. How well would we expect the 'winner' to perform?

The expected accuracy of the best participant will be: **80%**

With 1,048,576 participants we would expect one to get a perfect labeling

The chance of obtaining a very good labeling may be very high, even if the chance of obtaining a perfect labeling is very low.

Graph of Highest (Expected) Chance Accuracies



Perr = 0.5

Feature Selection: Highest Chance Accuracy

Can we use this theorem to provide a principled threshold for discriminative feature selection?

Feature Selection: Highest Chance Accuracy

- Yes, but…
- Theorem Assumptions
  - All the features (classifiers) are independent
  - All the features are noisy
  - The probability of a single classifier making a mistake is a Bernoulli trial with probability: p_{err}
- In Practice
  - Not all features are irrelevant
  - Often features are correlated

The theorem gives us a useful upper bound on expected accuracy of irrelevant features. But we'll need to relax this a bit to make it useful for feature selection. More later.

Feature Selection: Highest Chance Accuracy

- Suppose:
  - fMRI classification problem with 80,000 features
  - Train, then evaluate each feature on validation set of 40 examples
  - 2 classes

  - If a feature is irrelevant (noisy), expected accuracy of a single feature: 50% (20 out of 40 labels correct)
  - If all features irrelevant, expected accuracy of the "*best*" one is: 83% (33 correct)
  - If validation set had 20 examples instead: "best" noisy feature has 94% accuracy.

There is gap between the expected accuracy of an individual feature and the "best" one. Gap depends on number of examples and number of features.

Feature Selection: Multiplicity Gap

- Let X_i be the number of errors made by the i^{th} classifier:

$$X_1, X_2, \ldots, X_M \sim Binomial(N, p_{err})$$

- The classifier with the smallest number of errors:

$$X_{(1)} = min(X_1, X_2, \ldots, X_M)$$

- Define the *multiplicity gap*:

$$\mathcal{G}_{M,N} = \mathbb{E}[X] - \mathbb{E}[X_{(1)}]$$

Expected number of errors for an individual classifier

$$\mathbb{E}[X] = N \cdot p_{err}$$

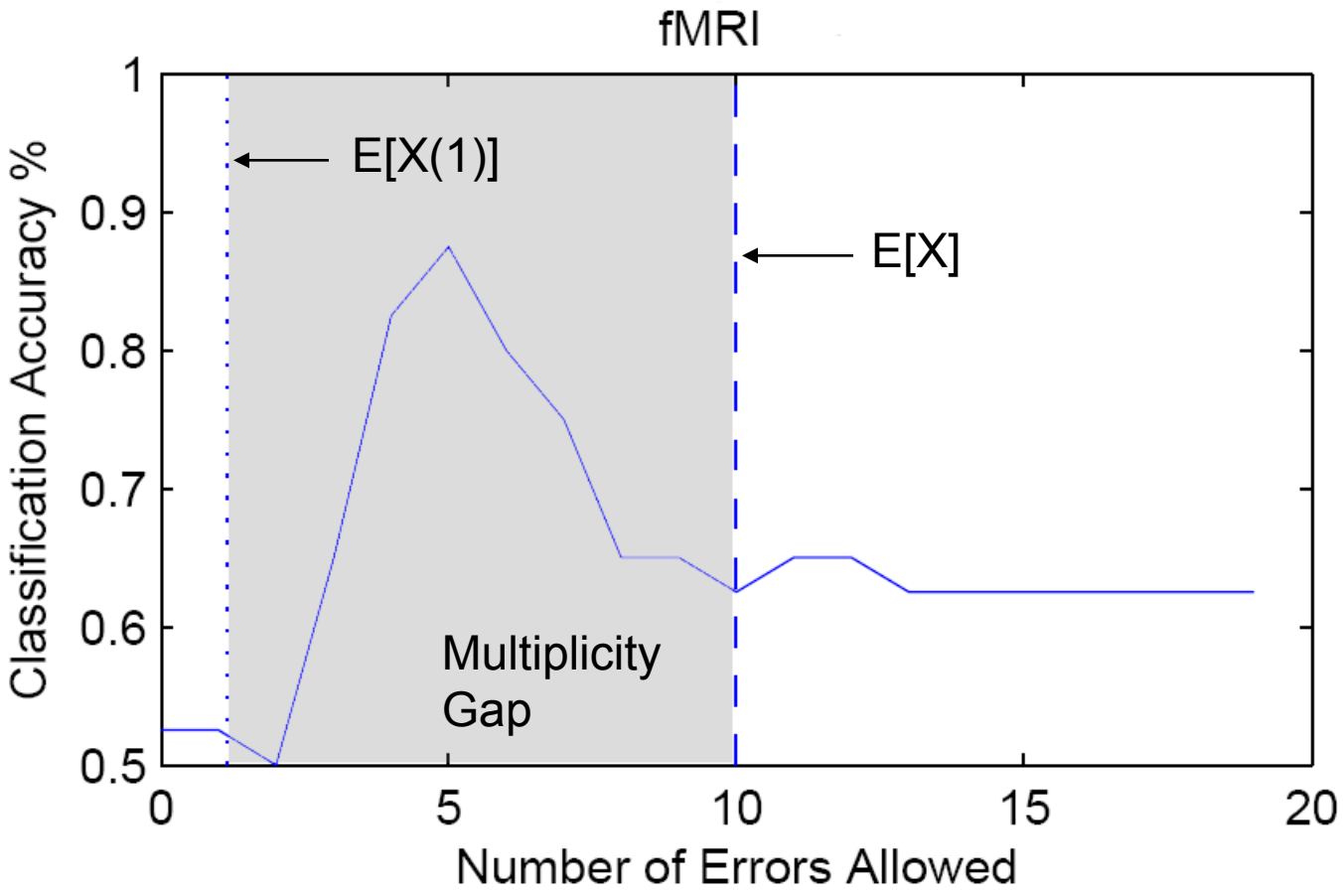
Expected number of errors for "best" irrelevant feature

Feature Selection: Highest Chance Accuracy

- Simple experiment:
  - fMRI cognitive state classification task: *Is the person viewing a picture or reading a sentence?*
  - 2 classes, 40 examples
  - Gaussian Naïve Bayes Classifier
  - Leave-One-Out-Cross Validation
  - Training:
    - Train on 19 examples
    - Evaluate on 20 validation examples. Choose features that made no more than XX errors.
    - Retrain on all 39 examples using just these features.
    - Test held out example
    - Repeat for each example
  - Repeat for each error threshold

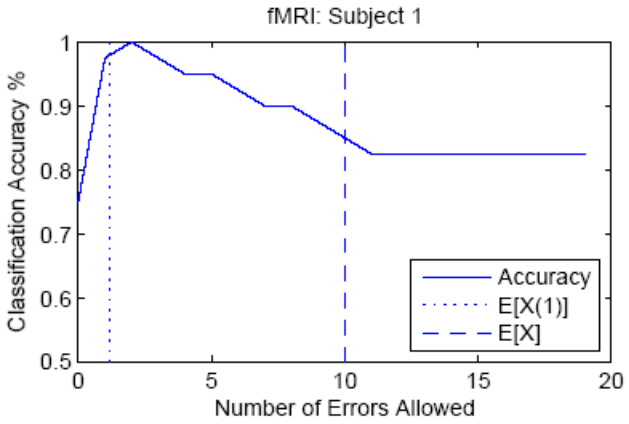
Feature Selection: Multiplicity Gap

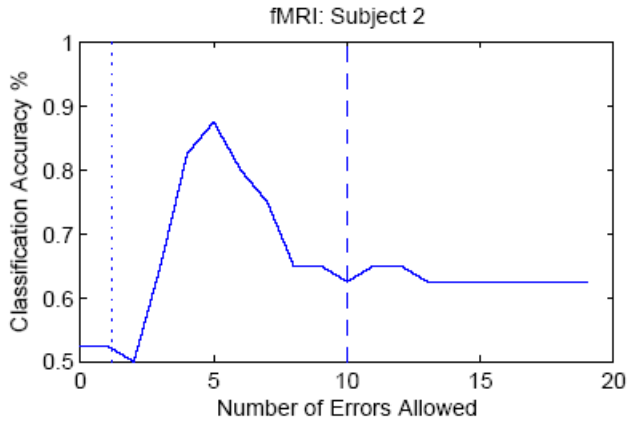
Conjecture: The optimal discriminative threshold will fall within the multiplicity gap

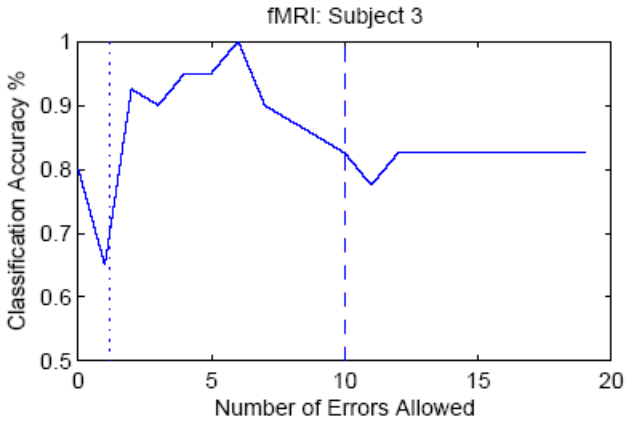


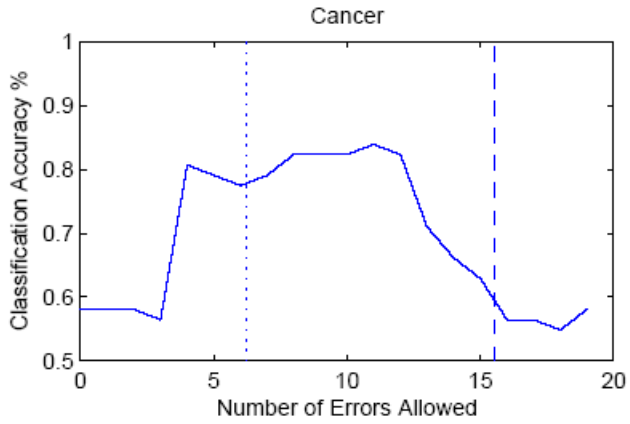
Feature Selection: Multiplicity Gap

Conjecture: The optimal discriminative threshold will fall within the multiplicity gap









Feature Selection: Multiplicity Gap Midpoint

- Empirically, we found the peak to fall in the multiplicity gap in all our experiments.
- To choose a good threshold:
  - Choose the midpoint of the two extremes
- Multiplicity Gap Midpoint
  - Feature selection heuristic that relaxes assumptions of theorem

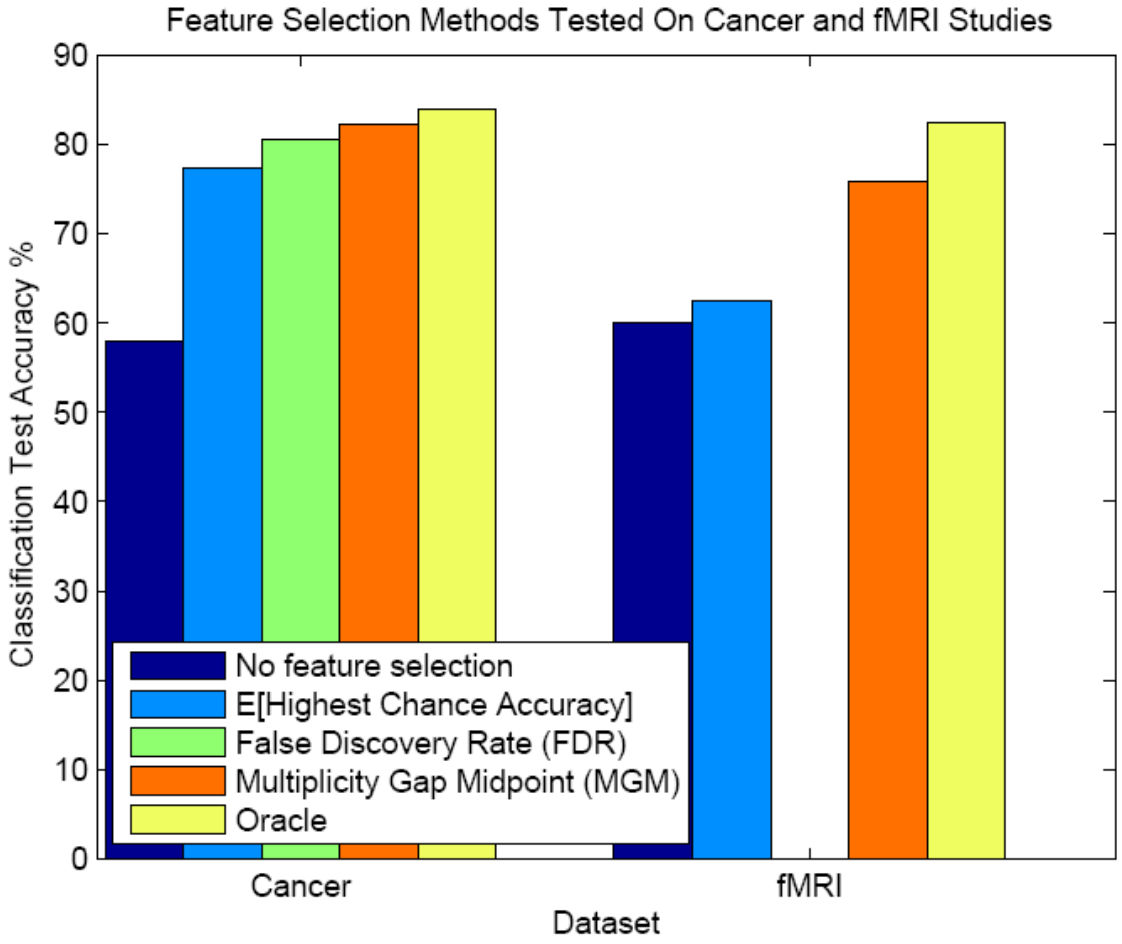
$$\tau_{MGM} = \frac{\left(\mathbb{E}\left[X\right] + \mathbb{E}\left[X_{(1)}\right]\right)}{2}$$

Feature Selection: Experimental Results

- Experiments:
  - fMRI classification task with 2 classes.
    - ~80,000 features, 40 examples
    - 13 subjects tested individually (results averaged)
  - Cancer detection using gene microarray
    - 2,000 features
    - 60 examples

- Gaussian Naïve Bayes
- Validation set is half of available training data
- Leave-one-out-cross-validation

Feature Selection: Experimental Results



Take Away Points

Highest Chance Accuracy Theorem

- Outputs intuitive accuracy number that defines a *natural significance threshold*

Multiplicity Gap Midpoint Heuristic:

- Relaxation provides useful feature selection threshold for sparse, high dimensional problems
- Elegant form which is computationally simple to calculate (1 line of Matlab)

See Palatucci and Carlson 2008:
On the Chance Accuracies of Large Collections of Classifiers

Thanks to:

Haikady Nagaraja

W.M. Keck Foundation

National Science Foundation

Yahoo!