Model Selection

1. Choose a class of models
2. Select the “best” member of that class
3. Use this model for all future predictions

Model Combination

- It is always possible to do at least as well as when using a single model
- Need to have a variety of classifiers
Model Selection

1. Choose a class of models
2. Select the “best” member of that class
3. Use this model for all future predictions

Model Combination

- It is always possible to do at least as well as when using a single model
- Need to have a variety of classifiers
- Populations not always stationary with time
- Classification performance may deteriorate

Example - changing priors
- Populations not always stationary with time
- Classification performance may deteriorate

Example - changing priors

![Graph showing changes in classification error rate over time]
Example - changing means

![Graph 1: Scatter plot showing data points with changing means.]

![Graph 2: Line graph showing error rate over time.]

Amber Tomas
Dynamic Mixture Models for Classification
The setting

- Use classifiers which output an estimate of the class probability distribution
- Have a batch of training data, assumed stationary
- Component classifiers are never retrained

A possible approach

- Use

\[ \hat{P}(Y = k|x; \hat{\beta}) = \sum_{i=1}^{M} \hat{\beta}_i \hat{p}^i(Y = k|x) \]

- This implies that \( \sum_{i=1}^{M} \hat{\beta}_i \) must equal one
The setting

- Use classifiers which output an estimate of the class probability distribution
- Have a batch of training data, assumed stationary
- Component classifiers are **never retrained**

A possible approach

- Use

\[
\hat{P}(Y = k | x; \hat{\beta}) = \sum_{i=1}^{M} \hat{\beta_i} \hat{p^i}(Y = k | x)
\]

- This implies that \(\sum_{i=1}^{M} \hat{\beta_i}\) must equal one
My approach

- Use a logistic model

\[ P(Y = k|x; \beta) = \frac{\exp\{\beta^T \eta_k(x)\}}{1 + \sum_{k=2}^{K} \exp\{\beta^T \eta_k(x)\}}, \quad k = 2, \ldots, K \]

- \( \eta_k(x) = (\eta_{1k}(x), \eta_{2k}(x), \ldots, \eta_{Mk}(x)) \) is user defined
- Forces the probabilities to sum to one
- The support of the posterior distribution of \( \beta \) is not constrained
Specification of $\eta_k, k = 1, 2, \ldots, K$

1. $\eta_k(x) = (\hat{p}^1(k|x), \ldots, \hat{p}^M(k|x))$

$$\implies \log\left(\frac{P(Y = k|x; \beta)}{P(Y = 1|x; \beta)}\right) = \sum_{i=1}^{M} \beta_i \hat{p}^i(k|x)$$

2. $\eta_k(x) = \left(\log\left(\frac{\hat{p}^1(k|x)}{\hat{p}^1(1|x)}\right), \ldots, \log\left(\frac{\hat{p}^M(k|x)}{\hat{p}^M(1|x)}\right)\right)$

$$\implies \log\left(\frac{P(Y = k|x; \beta)}{P(Y = 1|x; \beta)}\right) = \sum_{i=1}^{M} \beta_i \log\left(\frac{\hat{p}^i(k|x)}{\hat{p}^1(1|x)}\right)$$

Our choice determines how we should interpret the coefficients $\beta_i, i = 1, 2, \ldots, M$. 
Specification of $\eta_k, k = 1, 2, \ldots, K$

1. $\eta_k(x) = (\hat{p}^1(k|x), \ldots, \hat{p}^M(k|x))$

$$\Rightarrow \log \left( \frac{P(Y = k|x; \beta)}{P(Y = 1|x; \beta)} \right) = \sum_{i=1}^{M} \beta_i \hat{p}^i(k|x)$$

2. $\eta_k(x) = \left( \log \left( \frac{\hat{p}^1(k|x)}{\hat{p}^1(1|x)} \right), \ldots, \log \left( \frac{\hat{p}^M(k|x)}{\hat{p}^M(1|x)} \right) \right)$

$$\Rightarrow \log \left( \frac{P(Y = k|x; \beta)}{P(Y = 1|x; \beta)} \right) = \sum_{i=1}^{M} \beta_i \log \left( \frac{\hat{p}^i(k|x)}{\hat{p}^1(1|x)} \right)$$

Our choice determines how we should interpret the coefficients $\beta_i, i = 1, 2, \ldots, M$. 
Updating the parameters

- Prior distribution: $p(\beta)$
- Model: $\beta_t = \beta_{t-1} + \omega_t$, $\omega_t \sim N(0, V_t)$
- At each time point $t$,
  1. Observe feature vector $x_t$
  2. Calculate estimates $\hat{p}(Y_t = k|x_t)$, $k = 1, 2, \ldots, K$ and make a prediction
  3. Observe $y_t$ and update the posterior distribution of $\beta_t$ using

\[
p(y_t = k|x_t; \beta_t) = \frac{\exp\{\beta_t^T \eta_k(x_t)\}}{1 + \sum_{k=2}^{K} \exp\{\beta_t^T \eta_k(x_t)\}}
\]

- Approximations are necessary
Results: $\eta^i_k = \hat{p}^i(k|x)$
Results: $\eta^i_k = \log\{\hat{p}^i(k|x)/\hat{p}^i(1|x)\}$
Results: Drifting mean, $\pi^1 = 0.5$
Results: Drifting mean, $\pi^1 = 0.7$
Results: Trace of $\hat{\beta}_i, i = 1, 2, \ldots, 5$
There are several “tuning parameters” to be determined, including:

- The form of $\eta(x)$
- The evolution variance of $\beta_t = \beta_{t-1} + \omega_t$
- The number of component classifiers to use
- How we choose and train the component classifiers

Future work will be focused on

- Investigation
- Understanding
There are several “tuning parameters” to be determined, including

- The form of $\eta(x)$
- The evolution variance of $\beta_t = \beta_{t-1} + \omega_t$
- The number of component classifiers to use
- How we choose and train the component classifiers

Future work will be focused on

- Investigation
- Understanding