Information Theory
and
Feature Selection

(Joint Informativeness and Tractability)

Leonidas Lefakis

Zalando Research Labs

- Construction

$$X_1, \ldots, X_D \to f_1(X_1, \ldots, X_D), \ldots, f_k(X_1, \ldots, X_D)$$

- Construction

$$X_1, \ldots, X_D \to f_1(X_1, \ldots, X_D), \ldots, f_k(X_1, \ldots, X_D)$$

Examples

  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Autoencoders (Neural Networks)

- Selection

$$X_1, \ldots, X_D \to X_{s_1}, \ldots, X_{s_k}$$

  Approaches

  - Wrappers
  - Embedded methods
  - Filters

- Selection

$$X_1, \ldots, X_D \rightarrow X_{s_1}, \ldots, X_{s_k}$$

  - Wrappers

    Features are selected relative to the performance of a specific predictor.

    Example RFE-SVM.

- Selection

$$X_1, \ldots, X_D \rightarrow X_{s_1}, \ldots, X_{s_k}$$

  - Embedded Methods

    Features are selected internally while optimizing the predictor.

    Example Decision Trees.

- Selection

$$X_1, \ldots, X_D \to X_{s_1}, \ldots, X_{s_k}$$

  - Filters

    Features are assessed based on some goodness-of-fit function Φ that is classifier agnostic.

    Example Correlation.

Information Theory

- Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

- Joint Entropy

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

- Conditional Entropy

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

Information Theory

- Relative Entropy (Kullback-Leibler Divergence)

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- Mutual Information

$$I(X;Y) = \underbrace{\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy}_{D_{KL}(p(x,y)||p(x)p(y))}$$

Mutual Information

$$I(X;Y) = \underbrace{\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy}_{D_{KL}(p(x,y)||p(x)p(y))}$$

$$I(X;Y) = H(Y) - H(Y|X)$$

Reduction in uncertainty of Y if X is known

$$X \perp\!\!\!\perp Y \Rightarrow I(X;Y) = 0$$

$$Y = f(X) \Rightarrow I(X;Y) = H(Y)$$

We will look into feature selection in the context of classification. Given features

$$\{X_1, \ldots, X_D\} \in \mathbb{R}$$

and a class variable Y, we wish to select a subset S of size $K << D$ such that a predictor

$$f : \mathbb{R}^K \to Y$$

trained in this projected subspace generalizes well.

Entropy & Prediction

Given

$$X \times Y \in \mathbb{R}^D \times \{1 \dots C\}, \quad f(X) = \hat{Y}$$

We define the error variable

$$E = \begin{cases} 0 & \text{if } \hat{Y} = Y \\ 1 & \text{if } \hat{Y} \neq Y \end{cases}$$

Entropy & Prediction

$$
\begin{aligned}
H(E, Y|\hat{Y}) &\overset{(1)}{=} H(Y|\hat{Y}) + \underbrace{H(E|Y, \hat{Y})}_{=0} \\
&\overset{(1)}{=} \underbrace{H(E|\hat{Y})}_{\leq H(E)} + H(Y|E, \hat{Y})
\end{aligned}
$$

$$
H(A, B) = H(A) + H(B|A) \tag{1}
$$

Entropy & Prediction

$$
\begin{aligned}
H(Y|\hat{Y}) &= H(E, Y|\hat{Y}) \\
&\leq H(E) + H(Y|E, \hat{Y}) \\
&\leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)
\end{aligned}
$$

$$H(E) = H\left(B(1, P_e)\right) \leq H\left(B(1, \frac{1}{2})\right) = 1$$

$$H(Y|E, \hat{Y}) = (1 - P_e)\underbrace{H(Y|E = 0, \hat{Y})}_{=0} + P_e \underbrace{H(Y|E = 1, \hat{Y})}_{\leq \log(|\mathcal{Y}|\text{-}1)}$$

Fano's Inequality

zalando

$$H(Y|\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\overset{(2)}{\Longrightarrow} H(Y) - I(Y;\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\Longrightarrow P_e \geq \frac{H(Y) - I(Y;\hat{Y}) - 1}{\log(|\mathcal{Y}|\text{-}1)}$$

$$\overset{3}{\Longrightarrow} P_e \geq \frac{H(Y) - 1 - I(X;Y)}{\log(|\mathcal{Y}|\text{-}1)}$$

$$I(A;B) = H(A) - H(A|B) \tag{2}$$

$$Y \to X \to Z \Longrightarrow I(Y;X) \geq I(Y;Z) \tag{3}$$

Fano's Inequality

$$H(Y|\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\overset{(2)}{\implies} H(Y) - I(Y;\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\implies P_e \geq \frac{H(Y) - I(Y;\hat{Y}) - 1}{\log(|\mathcal{Y}|\text{-}1)}$$

$$\overset{3}{\implies} P_e \geq \frac{H(Y) - 1 - I(X;Y)}{\log(|\mathcal{Y}|\text{-}1)}$$

$$I(A;B) = H(A) - H(A|B) \tag{2}$$

$$Y \to X \to Z \implies I(Y;X) \geq I(Y;Z) \tag{3}$$

Data Processing Inequality

$$H(Y|\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\overset{(2)}{\Longrightarrow} H(Y) - I(Y;\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\Longrightarrow P_e \geq \frac{H(Y) - I(Y;\hat{Y}) - 1}{\log(|\mathcal{Y}|\text{-}1)}$$

$$\overset{3}{\Longrightarrow} P_e \geq \frac{H(Y) - 1 - I(X;Y)}{\log(|\mathcal{Y}|\text{-}1)}$$

$$I(A;B) = H(A) - H(A|B) \tag{2}$$

$$Y \to X \to Z \Longrightarrow I(Y;X) \geq I(Y;Z) \tag{3}$$

Fano's Inequality

zalando

$$H(Y|\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\overset{(2)}{\Longrightarrow} H(Y) - I(Y;\hat{Y}) \leq 1 + P_e \log(|\mathcal{Y}|\text{-}1)$$

$$\Longrightarrow P_e \geq \frac{H(Y) - I(Y;\hat{Y}) - 1}{\log(|\mathcal{Y}|\text{-}1)}$$

$$\overset{3}{\Longrightarrow} P_e \geq \frac{H(Y) - 1 - I(X;Y)}{\log(|\mathcal{Y}|\text{-}1)}$$

$$I(A;B) = H(A) - H(A|B) \tag{2}$$

$$Y \to X \to Z \Longrightarrow I(Y;X) \geq I(Y;Z) \tag{3}$$

Objective

$$S = \underset{S', |S'|=K}{\operatorname{argmax}} I(X_{S'(1)}, X_{S'(2)}, \ldots, X_{S'(K)}; Y)$$

Objective

$$S = \underset{S', |S'|=K}{\operatorname{argmax}} I(X_{S'(1)}, X_{S'(2)}, \ldots, X_{S'(K)}; Y)$$

NP-HARD

Feature Selection

Naive

$$S = \underset{S', |S'|=K}{\operatorname{argmax}} \sum_{k=1}^{K} I(X_{S'(k)}; Y)$$

- Considers the Relevance of each variable individually
- Does not consider Redundancy
- Does not consider Joint Informativeness

Feature Selection

Two Popular Solutions

- **mRMR** : Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, H.Peng et al.
- **CMIM** : Fast Binary Feature Selection with Conditional Mutual Information, F.Fleuret

mRMR

Relevance

$$D(S, Y) = \frac{1}{|S|} \sum_{X_d \in S} I(X_d; Y)$$

Redundancy

$$R(S) = \frac{1}{|S|^2} \sum_{X_d \in S} \sum_{X_l \in S} I(X_d; X_l)$$

mRMR

$$\Phi(S, Y) = D(S, Y) - R(S)$$

Forward Selection

$S_0 = \emptyset$
for $k = 1 \ldots K$ **do**
  $z^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \Phi(S', Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

mRMR

How do we calculate $I(\cdot;\cdot)$?

- Discretize
- Distributions approximated using Parzen windows

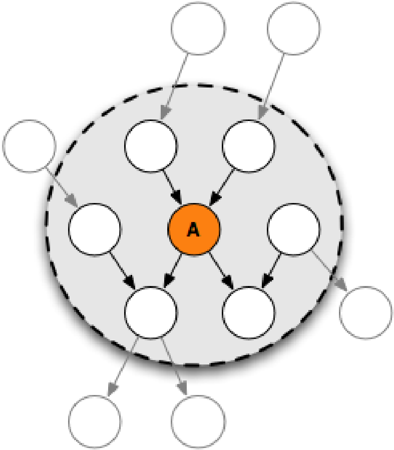
$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x^{(i)}, h)$$

$$\delta(dx, h) = \frac{1}{\sqrt{2\pi^D}} exp\left(-\frac{dx^T\Sigma^{-1}dx}{2h^2}\right)$$

- Parametric Model

CMIM (Binary Features)

Markov Blanket



Markov Blanket of variable A [1]

[1]Wiki Commons

For a set M of variables that form a Markov Blanket of X_i we have

$$p\left(F \setminus \{X_i, M\}, Y | X_i, M\right) = p\left(F \setminus \{X_i, M\}, Y | M\right)$$

$$I\left(F \setminus \{X_i, M\}, Y; X_i | \, M\right) = 0$$

CMIM (Binary Features)

Markov Blanket

For a set M of variables that form a Markov Blanket of X_i we have

$$p\left(F \setminus \{X_i, M\}, Y | X_i, M\right) = p\left(F \setminus \{X_i, M\}, Y | M\right)$$

$$I\left(F \setminus \{X_i, M\}, Y; X_i \middle| M\right) = 0$$

For Feature Selection $\rightarrow$ Too Strong

$$I\left(Y; X_i \middle| M\right) = 0$$

CMIM (Binary Features)

$$I(X_1, \ldots, X_K; Y) = H(Y) - \underbrace{H(Y|X_1, ..X_K)}_{\text{intractable}}$$

CMIM (Binary Features)

$$I(X_1, \ldots, X_K; Y) = H(Y) - \underbrace{H(Y|X_1, ..X_K)}_{\text{intractable}}$$

$$I(X_i; Y|X_j) = H(Y|X_j) - H(Y|X_i, X_j)$$

Distributions over triplets of variables

CMIM (Binary Features)

zalando

$$I(X_1, \ldots, X_K; Y) = H(Y) - \underbrace{H(Y|X_1, ..X_K)}_{\text{intractable}}$$

$$I(X_i; Y|X_j) = H(Y|X_j) - H(Y|X_i, X_j)$$

$$S_1 \leftarrow \underset{d}{\text{argmax}}\ \hat{I}(X_d; Y)$$

$$S_k \leftarrow \underset{d}{\text{argmax}}\{\ \min_{l<k} \hat{I}(Y; X_d|X_{S_l})\ \}$$

Distributions over triplets of variables

Fast CMIM

$$S_k \leftarrow \underset{d}{\operatorname{argmax}}\{ \underbrace{\min_{l<k} \hat{I}(Y; X_d | X_{S_l})}_{\text{Can only decrease}} \}$$

Fast CMIM

$$S_k \leftarrow \underset{d}{\operatorname{argmax}} \{ \underbrace{\min_{l<k} \hat{I}(Y; X_d | X_{S_l})}_{\text{Can only decrease}} \}$$

	1	2	3	4	5	6	7
1	**6**	**3**	**1**	**5**	**4**	**2**	**5**
2	**3**	**5**	?	**5**	**3**	?	**3**
3	**5**	**2**	?	**4**	?	?	?
4	**4**	?	?	**6**	?	?	?
5	**3**	?	?	**4**	?	?	?
ps[n]	3	2	1	4	3	2	3
m[n]	5	3	1	5	2	1	2

$$ps[n] = \min_{l<m(n)} \hat{I}(Y; X_d | X_{S_l})$$

Objective

$$S = \operatorname*{argmax}_{S', |S'|=K} I(X_{S'(1)}, X_{S'(1)}, \ldots, X_{S'(K)}; Y)$$

CMIM, mRMR (and others)

Compromise Joint Informativeness for Tractability

Problem

$$\{X_1, \ldots, X_D\} \in \mathbb{R}, Y \in \{1...C\}$$

$$S = \underset{S', |S'|=K}{\operatorname{argmax}} I(X_{S'(1)}, X_{S'(2)}, \ldots, X_{S'(K)}; Y)$$

Calculate $I(X; Y)$ using a joint law

Information Theory

zalando

- Differential Entropy

$$h(X) = -\int_X p(x) \log\left(p(x)\right) dx$$

- Relative Entropy (Kullback-Leibler Divergence)

$$D(p\|q) = \int_X p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

- Mutual Information

$$I\left(X;Y\right) = \int_X \int_Y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

Addressing Intractability

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y)\,\underbrace{H(X|Y=y)}_{\text{intractable}}
\end{aligned}
$$

Addressing Intractability

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y) \underbrace{H(X|Y=y)}_{\text{intractable}}$$

Parametric model $p_{X|Y}$

Addressing Intractability

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y)\underbrace{H(X|Y=y)}_{\text{intractable}}$$

Parametric model $p_{X|Y} = N(\mu_y, \Sigma_y)$

Addressing Intractability

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y)\underbrace{H(X|Y=y)}_{\text{tractable}}$$

Parametric model $p_{X|Y} = N(\mu_y, \Sigma_y)$

$$H(X|Y) = \frac{1}{2}\log(|\Sigma_y|) + \frac{n}{2}(\log 2\pi + 1)\,.$$

Maximum Entropy Distribution

Given

$$E(x) = \mu, E\left((x-\mu)(x-\mu)^T\right) = \Sigma$$

then the multivariate Normal

$$\vec{x} \sim N(\vec{\mu}, \Sigma)$$

is the Maximum Entropy Distribution

Addressing Intractability

zalando

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y) \underbrace{H(X|Y=y)}_{\text{tractable}}$$

Parametric model $p_{X|Y} = N(\mu_y, \Sigma_y)$

$$H(X|Y) = \frac{1}{2}\log(|\Sigma_y|) + \frac{n}{2}(\log 2\pi + 1).$$

Addressing Intractability

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{\color{red}{H(X)}}_{\text{intractable}} - \sum_{Y} P(Y=y) \underbrace{H(X|Y=y)}_{\text{tractable}}$$

$$H(X) = H\left(\sum_{y} \pi_y N(\mu_y, \Sigma_y)\right)$$

Addressing Intractability

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y)\,\underbrace{H(X|Y=y)}_{\text{tractable}}$$

$$H(X) = H\left(\sum_{y} \pi_y N(\mu_y, \Sigma_y)\right) \rightarrow \underbrace{\text{Entropy of Mixture of Gaussians}}_{\text{No Analytical Solution}}$$

Addressing Intractability

zalando

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \underbrace{H(X)}_{\text{intractable}} - \sum_{Y} P(Y=y)\,\underbrace{H(X|Y=y)}_{\text{tractable}}$$

$$H(X) = H\left(\sum_{y} \pi_y N(\mu_y, \Sigma_y)\right) \rightarrow \underbrace{\text{Entropy of Mixture of Gaussians}}_{\text{No Analytical Solution}}$$

Upper Bound or Approximate

$$H\left(\sum_{y} \pi_y N(\mu_y, \Sigma_y)\right)$$

A Normal Upper Bound

$$p_X = \sum_Y \pi_y p_{X|Y}$$

A Normal Upper Bound

$$p_X = \sum_Y \pi_y p_{X|Y}$$

$$p* \sim N(\mu^*, \Sigma^*) \stackrel{maxEnt}{\Longrightarrow} H(p_X) \leq H(p*)$$

A Normal Upper Bound

$$p_X = \sum_Y \pi_y p_{X|Y}$$

$$p* \sim N(\mu^*, \Sigma^*) \overset{maxEnt}{\Longrightarrow} H(p_X) \leq H(p*)$$

$$I(X;Y) \leq H(p*) - \sum_Y P_Y H(X|Y)$$

A Normal Upper Bound

$$p_X = \sum_Y \pi_y p_{X|Y}$$

$$p* \sim N(\mu^*, \Sigma^*) \overset{maxEnt}{\Longrightarrow} H(p_X) \leq H(p*)$$

$$I(X;Y) \leq H(p*) - \sum_Y P_Y H(X|Y)$$

$$I(X;Y) \leq H(Y) = -\sum_y p_y \log p_y$$

A Normal Upper Bound

$$p_X = \sum_Y \pi_y p_{X|Y}$$

$$I(X;Y) \leq H(p*) - \sum_Y P_Y H(X|Y)$$

$$I(X;Y) \leq H(Y) = -\sum_y p_y \log p_y$$

A Normal Upper Bound

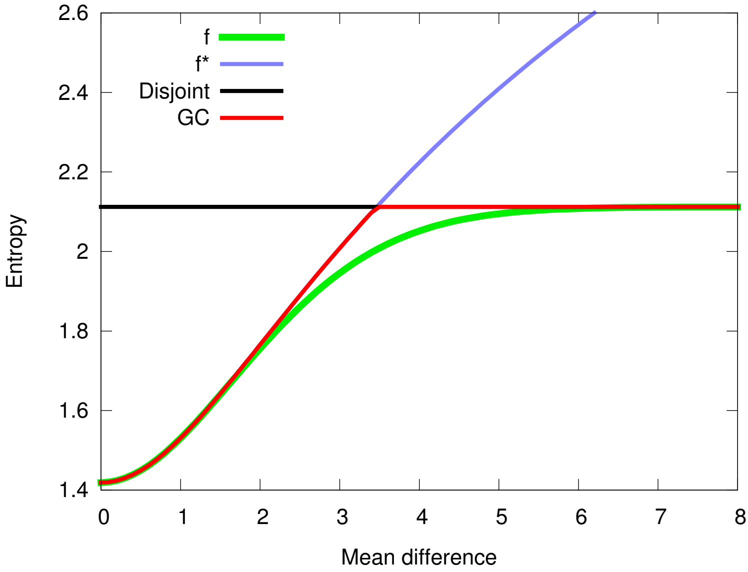
zalando

$$p_X = \sum_Y \pi_y p_{X|Y}$$

$$I(X;Y) \leq H(p*) - \sum_Y P_Y H(X|Y)$$

$$I(X;Y) \leq H(Y) = -\sum_y p_y \log p_y$$

$$I(X;Y) \leq \min\left(H(p^*), \sum_Y P_Y \left(H(X|Y) - \log(P_Y)\right)\right) - \sum_Y P_Y H(X|Y)$$


f
f*
Disjoint
GC
Entropy
2.6
2.4
2.2
2
1.8
1.6
1.4
0
1
2
3
4
5
6
7
8
Mean difference

An approximation

$$p_X = \sum_Y \pi_y p_{X|Y}$$

Under mild assumptions

$$\forall y, H(p^*) > H(p_{X|Y=y})$$

we can use an approximation to $I(X;Y)$

$$\tilde{I}(X;Y) = \sum_y \min(H(p^*), H(X|Y) - \log p_y) p_y - \sum_y H(X|Y) p_y$$

Feature Selection Criterium

Mutual Information Approximation

$$S = \underset{S', |S'|=K}{\operatorname{argmax}} \left(\tilde{I}(X_{S'(1)}, X_{S'(2)}, \ldots, X_{S'(K)}; Y) \right)$$

Forward Selection

$S_0 = \emptyset$
for $k = 1 \dots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \hat{I}(S'; Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

$$\hat{I}(S'; Y) \propto \sum_y \min\left(\log(|\Sigma^*|), \log(|\Sigma_y|) - \log p_y\right) p_y - \sum_y \log(|\Sigma_y|) p_y$$

Complexity

At iteration k we need to calculate $\forall j \in F \setminus S_{k-1}$

$$\left|\Sigma_{S_{k-1} \cup X_j}\right|$$

Complexity

At iteration k we need to calculate $\forall j \in F \setminus S_{k-1}$

$$\left|\Sigma_{S_{k-1} \cup x_j}\right|$$

The cost of calculating each determinant is $O(k^3)$

Forward Selection

$S_0 = \emptyset$
for $k = 1 \ldots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \tilde{I}(S'; Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^4)$

Forward Selection

zalando

$S_0 = \emptyset$
for $k = 1 \dots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \tilde{I}(S'; Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^4)$

However...

Complexity

$$\Sigma_{S_{k-1} \cup X_j} = \begin{bmatrix} \Sigma_{S_{k-1}} & \Sigma_{j,S_{k-1}} \\ \Sigma_{j,S_{k-1}}^T & \sigma_j^2 \end{bmatrix}$$

Complexity

$$\Sigma_{S_{k-1} \cup X_j} = \begin{bmatrix} \Sigma_{S_{k-1}} & \Sigma_{j,S_{k-1}} \\ \Sigma_{j,S_{k-1}}^T & \sigma_j^2 \end{bmatrix}$$

We can exploit the matrix determinant lemma (twice)

$$\left|\Sigma + uv^T\right| = \left(1 + v^T \Sigma^{-1} u\right) |\Sigma|$$

To compute each determinant in $O(n^2)$

Forward Selection

$S_0 = \emptyset$
for $k = 1 \ldots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \tilde{I}(S'; Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^3)$

Forward Selection

$S_0 = \emptyset$
for $k = 1 \ldots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \tilde{I}(S'; Y)$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^3)$

Faster?

Faster?

zalando

$$I(X, Z; Y) = I(Z; Y) + I(X; Y \mid Z)$$

$$I(S_k'; Y) = I(X_j; Y \mid S_{k-1}) + \underbrace{I(S_{k-1}; Y)}_{\text{common}}$$

Faster?

$$I(X, Z; Y) = I(Z; Y) + I(X; Y \mid Z)$$

$$I(S_k'; Y) = I(X_j; Y \mid S_{k-1}) + \underbrace{\cancel{I(S_{k-1}; Y)}}_{\text{common}}$$

$$\underset{X_j \in F \setminus S_{k-1}}{\operatorname{argmax}} \; I(X_j; Y \mid S_{k-1}) = \underset{X_j \in F \setminus S_{k-1}}{\operatorname{argmax}} \; (H(X_j \mid S_{k-1}) - H(X_j \mid Y, S_{k-1}))$$

Conditional Entropy

$$H(X_j \mid S_{k-1}) = \int_{\mathbb{R}^{|S_{k-1}|}} H(X_j \mid S_{k-1} = s)\mu_{S_{k-1}}(s)ds$$
$$= \frac{1}{2}\log\sigma^2_{j|S_{k-1}} + \frac{1}{2}(\log 2\pi + 1)$$

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{j,S_{k-1}}\Sigma^{-1}_{S_{k-1}}\Sigma_{j,S_{k-1}}.$$

Updating $\sigma^2_{j|S_{k-1}}$

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{j,S_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{j,S_{k-1}}$$

Assume X_i was chosen at iteration $k-1$

$$\begin{aligned}
\Sigma_{S_{k-1}} &= \begin{bmatrix} \Sigma_{S_{k-2}} & \Sigma_{i,S_{k-2}} \\ \Sigma^T_{i,S_{k-2}} & \sigma^2_i \end{bmatrix} \\
&= \begin{bmatrix} \Sigma_{S_{k-2}} & 0_{k-2} \\ 0^T_{n-2} & \sigma^2_i \end{bmatrix} + \boldsymbol{e}_{n+1} \Sigma^T_{i,S_{k-2}} + \Sigma_{i,S_{k-2}} \boldsymbol{e}^T_{n+1}
\end{aligned}$$

Updating $\Sigma_{S_{k-1}}^{-1}$

From Sherman-Morrison formula

$$\left(\Sigma + uv^T\right)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}uv^T\Sigma^{-1}}{1 + v^T\Sigma^{-1}u}$$

Updating $\Sigma^{-1}_{S_{k-1}}$

From Sherman-Morrison formula

$$\left(\Sigma + uv^T\right)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}uv^T\Sigma^{-1}}{1 + v^T\Sigma^{-1}u}$$

$$\Sigma^{-1}_{S_{n-1}} = \begin{bmatrix} \Sigma^{-1}_{S_{n-2}} & -\frac{1}{\beta\sigma_i^2}u \\ -\frac{1}{\beta\sigma_i^2}u^T & \frac{1}{\beta\sigma_i^2} \end{bmatrix} + \frac{1}{\beta\sigma_i^2}\begin{bmatrix} u \\ 0 \end{bmatrix}\begin{bmatrix} u^T & 0 \end{bmatrix}$$

Updating $\Sigma_{S_{k-1}}^{-1}$

zalando

$$\Sigma_{S_{n-1}}^{-1} = \begin{bmatrix} \Sigma_{S_{n-2}}^{-1} & -\frac{1}{\beta\sigma_i^2} u \\ -\frac{1}{\beta\sigma_i^2} u^T & \frac{1}{\beta\sigma_i^2} \end{bmatrix} + \frac{1}{\beta\sigma_i^2} \begin{bmatrix} u \\ 0 \end{bmatrix} \begin{bmatrix} u^T & 0 \end{bmatrix}$$

$$\begin{aligned} \sigma_{j|S_{n-1}}^2 = \sigma_j^2 - & \overbrace{\Sigma_{j,S_{n-2}}^T \Sigma_{S_{n-2}}^{-1} \Sigma_{j,S_{n-2}}}^{\text{Previous Round}} & \in O(n^2) \\ & + \frac{\sigma_{ji}^2}{\beta\sigma_i^2} u^T \Sigma_{j,S_{n-2}} - \Sigma_{j,S_{n-1}}^T \begin{bmatrix} -\frac{1}{\beta\sigma_i^2} u \\ \frac{1}{\beta\sigma_i^2} \end{bmatrix} \sigma_{ji}^2 & \in O(n) \\ & - \frac{1}{\beta\sigma_i^2} \left(\Sigma_{j,S_{n-1}}^T \begin{bmatrix} u \\ 0 \end{bmatrix} \right) \left(\begin{bmatrix} u^T & 0 \end{bmatrix} \Sigma_{jS_{n-1}} \right) & \in O(n) \end{aligned}$$

Faster!

zalando

for $k = 1 \dots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \hat{I}(S')$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^2)$

Faster!

for $k = 1 \dots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \hat{I}(S')$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Overall Complexity $O(|Y||F|K^2)$

Even Faster?

Even Faster?

Main bottleneck $O(k)$

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

Even Faster?

Main bottleneck $O(k)$

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

Skip non-promising features

Forward Selection

zalando

$S_0 = \emptyset$
for $k = 1 \dots K$ **do**
  $s^* = 0$
  for $X_j \in F \setminus S_{k-1}$ **do**
    $S^{'} \leftarrow S_{k-1} \cup X_j$
    $z \leftarrow \hat{I}(S')$
    if $z > z^*$ **then**
      $s^* \leftarrow s$
      $S^* \leftarrow S'$
    end if
  end for
  $S_i \leftarrow S^*$
end for
return S_K

Cheap ($O(1)$) score c:

if $c < z^*$ then $z < z^*$

Forward Selection

$\ldots$

$z^* = 0$

for $X_j \in F \setminus S_{k-1}$ **do**

  $S' \leftarrow S_{k-1} \cup X_j$

  $c \leftarrow ?$

  if $c > z^*$ **then**

    $z \leftarrow \hat{I}(S')$

    if $z > z^*$ **then**

      $s^* \leftarrow s$

      $S^* \leftarrow S'$

    end if

  end if

end for

$\ldots$

Cheap ($O(1)$) score c:

if $c < z^*$ then $z < z^*$

An $O(1)$ Bound

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

$$\Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} = \Sigma^T_{jS_{k-1}} U \Lambda U^T \Sigma_{jS_{k-1}}$$

$$\|\Sigma^T_{jS_{k-1}}\|^2_2 \max_i \lambda_i \geq \Sigma^T_{jS_{n-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} \geq \underbrace{\|\Sigma^T_{jS_{k-1}}\|^2_2 \min_i \lambda_i}_{O(1)}$$

An $O(1)$ Bound

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

$$\Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} = \Sigma^T_{jS_{k-1}} U \Lambda U^T \Sigma_{jS_{k-1}}$$

$$\|\Sigma^T_{jS_{k-1}}\|_2^2 \max_i \lambda_i \geq \Sigma^T_{jS_{n-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} \geq \underbrace{\|\Sigma^T_{jS_{k-1}}\|_2^2 \min_i \lambda_i}_{O(1)}$$

An $O(1)$ Bound

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

$$\Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} = \Sigma^T_{jS_{k-1}} U \Lambda U^T \Sigma_{jS_{k-1}}$$

$$\|\Sigma^T_{jS_{k-1}}\|^2_2 \max_i \lambda_i \geq \Sigma^T_{jS_{n-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} \geq \underbrace{\|\Sigma^T_{jS_{k-1}}\|^2_2 \min_i \lambda_i}_{O(1)}$$

An $O(1)$ Bound

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

$$\Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} = \Sigma^T_{jS_{k-1}} U \Lambda U^T \Sigma_{jS_{k-1}}$$

$$\|\Sigma^T_{jS_{k-1}}\|^2_2 \max_i \lambda_i \geq \Sigma^T_{jS_{n-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} \geq \underbrace{\|\Sigma^T_{jS_{k-1}}\|^2_2 \min_i \lambda_i}_{O(1)}$$

An $O(1)$ Bound

$$\sigma^2_{j|S_{k-1}} = \sigma^2_j - \Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}}$$

$$\Sigma^T_{jS_{k-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} = \Sigma^T_{jS_{k-1}} U \Lambda U^T \Sigma_{jS_{k-1}}$$

$$\|\Sigma^T_{jS_{k-1}}\|^2_2 \max_i \lambda_i \geq \Sigma^T_{jS_{n-1}} \Sigma^{-1}_{S_{k-1}} \Sigma_{jS_{k-1}} \geq \underbrace{\|\Sigma^T_{jS_{k-1}}\|^2_2 \min_i \lambda_i}_{O(1)}$$

EigenSystem Update Problem

Given U^n, Λ^n such that

$$\Sigma^n = {U^n}^T \Lambda^n U^n$$

EigenSystem Update Problem

Given U^n, Λ^n such that

$$\Sigma^n = {U^n}^T \Lambda^n U^n$$

Find U^{n+1}, Λ^{n+1}

$$\Sigma^{n+1} = \begin{bmatrix} \Sigma^n & \boldsymbol{v} \\ \boldsymbol{v}^T & 1 \end{bmatrix} = {U^{n+1}}^T \Lambda^{n+1} U^{n+1}$$

assume $\Sigma^{n+1} \in S^{n+1}_{++}$

EigenSystem Update

$$U^n = \left[\begin{array}{cccc} \boldsymbol{u}_1^n & \boldsymbol{u}_2^n & \dots & \boldsymbol{u}_n^n \end{array}\right], \quad \Lambda^n = \begin{bmatrix} \lambda_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^n \end{bmatrix}$$

EigenSystem Update

$$U^n = \left[\begin{array}{cccc} \boldsymbol{u}_1^n & \boldsymbol{u}_2^n & \dots & \boldsymbol{u}_n^n \end{array} \right], \quad \Lambda^n = \begin{bmatrix} \lambda_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^n \end{bmatrix}$$

$$\begin{bmatrix} \Sigma^n & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i^n \\ 0 \end{bmatrix} = \lambda_i^n \underbrace{\begin{bmatrix} \boldsymbol{u}_i^n \\ 0 \end{bmatrix}}_{\boldsymbol{u}_i^{\prime n}}$$

EigenSystem Update

$$U^n = \left[\begin{array}{cccc} \boldsymbol{u}_1^n & \boldsymbol{u}_2^n & \dots & \boldsymbol{u}_n^n \end{array} \right], \quad \Lambda^n = \begin{bmatrix} \lambda_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^n \end{bmatrix}$$

$$\begin{bmatrix} \Sigma^n & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i^n \\ 0 \end{bmatrix} = \lambda_i^n \underbrace{\begin{bmatrix} \boldsymbol{u}_i^n \\ 0 \end{bmatrix}}_{\boldsymbol{u}_i'^n}$$

$$\begin{bmatrix} \Sigma^n & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$$

EigenSystem Update

$$\underbrace{\begin{bmatrix} \Sigma^n & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}}_{\Sigma'} = \begin{bmatrix} \boldsymbol{u}_1'^{n\,T} \\ \vdots \\ \boldsymbol{e}^{n+1\,T} \end{bmatrix} \underbrace{\begin{bmatrix} \lambda_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}}_{\Lambda'} \underbrace{\begin{bmatrix} \boldsymbol{u}_1'^{n} & \dots & \boldsymbol{e}^{n+1} \end{bmatrix}}_{U'}$$

$$\Sigma^{n+1} = \Sigma' + \boldsymbol{e}_{n+1} \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix}^T + \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix} \boldsymbol{e}_{n+1}^T$$

$$\left(\Sigma' + \boldsymbol{e}_{n+1} \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix}^T + \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix} \boldsymbol{e}_{n+1}^T \right) u^{n+1} = \lambda^{n+1} u^{n+1}$$

EigenSystem Update

$$\left(\Sigma' + \boldsymbol{e}_{n+1}\begin{bmatrix}\boldsymbol{v}\\0\end{bmatrix}^T + \begin{bmatrix}\boldsymbol{v}\\0\end{bmatrix}\boldsymbol{e}_{n+1}^T\right)u^{n+1} = \lambda^{n+1}u^{n+1}$$

$$U'^T\left(\Sigma' + \boldsymbol{e}_{n+1}\begin{bmatrix}\boldsymbol{v}\\0\end{bmatrix}^T + \begin{bmatrix}\boldsymbol{v}\\0\end{bmatrix}\boldsymbol{e}_{n+1}^T\right)\underbrace{U'U'^T}_{U'U'^T=I}u^{n+1} = \lambda^{n+1}U'^T u^{n+1}$$

$$\stackrel{((4))}{\Longrightarrow}\left(\Lambda' + \boldsymbol{e}_{n+1}\boldsymbol{q}^T + \boldsymbol{q}\boldsymbol{e}_{n+1}^T\right)U'^T u^{n+1} = \lambda^{n+1}U'^T u^{n+1}$$

$$U'^T\Sigma' U' = \Lambda' \tag{4}$$

EigenSystem Update

$$\underbrace{\left(\Lambda' + \boldsymbol{e}_{n+1}\boldsymbol{q}^T + \boldsymbol{q}\boldsymbol{e}_{n+1}^T\right)}_{\Sigma''} U'^T u^{n+1} = \lambda^{n+1} U'^T u^{n+1}$$

$\rightarrow$ Σ^{n+1} and Σ'' share eigenvalues.

$\rightarrow$ $U^{n+1} = U'U''$

EigenSystem Update

$$|\Sigma'' - \lambda I| = \prod_j (\lambda_j' - \lambda) + \sum_{i<n+1} -q_i^2 \prod_{j \neq i, j<n+1} (\lambda_j' - \lambda)$$

$$f(\lambda) = \lambda'_{n+1} - \lambda + \sum_i \frac{-q_i^2}{(\lambda'_i - \lambda)}.$$

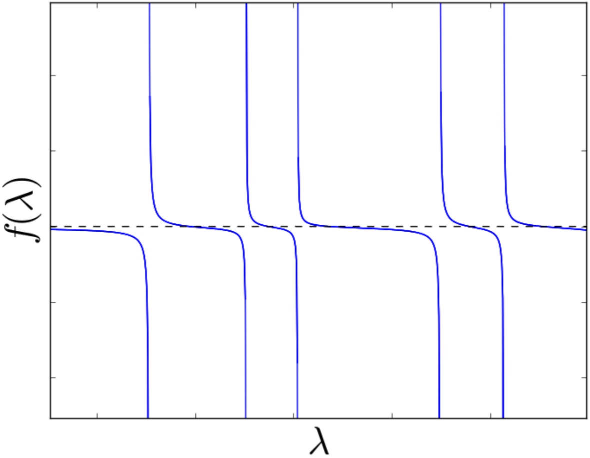
EigenSystem Update

$$|\Sigma'' - \lambda I| = \prod_j (\lambda'_j - \lambda) + \sum_{i<n+1} -q_i^2 \prod_{j\neq i, j<n+1} (\lambda'_j - \lambda)$$

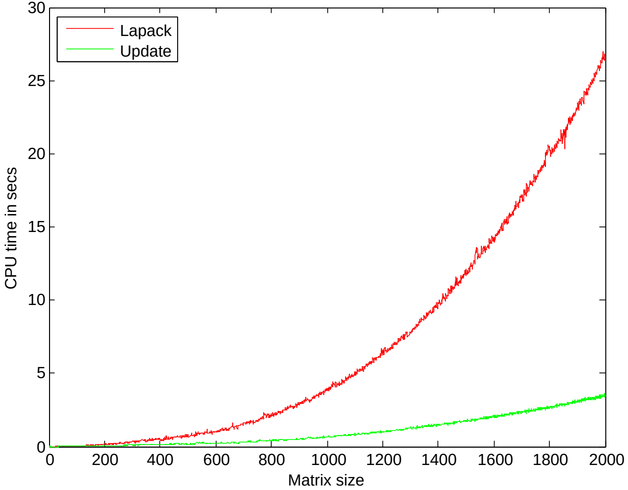
$$f(\lambda) = \lambda'_{n+1} - \lambda + \sum_i \frac{-q_i^2}{(\lambda'_i - \lambda)}.$$

$$\forall i, \lim_{\lambda \xrightarrow{>} \lambda'_i} f(\lambda) = +\infty, \lim_{\lambda \xrightarrow{<} \lambda'_i} f(\lambda) = -\infty$$

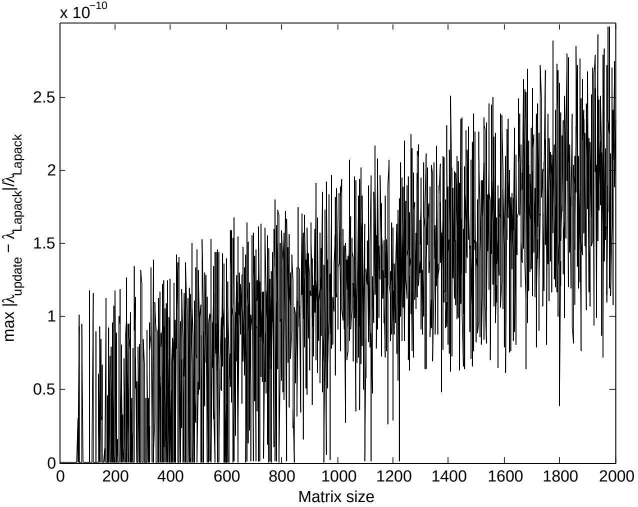
$$\frac{\partial f(\lambda)}{\partial \lambda} = -1 + \sum_i \frac{-q_i^2}{(\lambda'_i - \lambda)^2} \leq 0$$



$f(\lambda)$
λ



Comparison between scratch and update



Numerical stability (eigenvalues)

Set to go ...

Now that all the machinery is in place, How did we do?

Nice Results

SVMLin	CIFAR 10	CIFAR 50	CIFAR 100	STL 10	STL 50	STL 100	INRIA 10	INRIA 50	INRIA 100
Fisher	25.19	39.47	48.12	26.09	34.63	38.02	92.55	94.03	94.68
FCBF	**33.65**	47.77	54.97	31.74	38.11	40.66	**94.14**	**96.03**	96.03
MRMR	27.94	37.78	43.63	28.26	31.16	33.12	86.03	86.77	86.72
SBMLR	30.43	**51.41**	**56.81**	**32.29**	43.29	47.15	85.92	88.57	88.64
tTest	25.69	40.17	45.12	26.72	36.23	39.14	80.01	87.64	89.23
InfoGain	24.79	37.98	47.37	27.17	33.70	37.84	92.35	93.75	94.68
Spec. Clus.	17.19	32.78	42.6	18.91	32.65	38.24	**92.67**	93.64	94.44
RelieFF	24.56	38.17	46.51	29.16	38.05	42.94	90.99	**95.97**	**96.36**
CFS	31.49	42.17	51.70	28.63	38.54	41.88	88.64	96.11	**97.53**
CMTF	21.10	40.39	47.71	27.61	38.99	42.32	79.09	89.49	93.01
BAHSIC	-	-	-	28.95	39.05	45.49	78.54	89.77	91.96
GC.E	32.45	50.15	55.06	31.20	43.31	**49.75**	87.73	91.96	93.13
GC. MI	**36.47**	**51.44**	**55.39**	**32.50**	**44.15**	**48.88**	89.76	**95.71**	**96.45**
GKL.E	**37.51**	**52.11**	**56.41**	**33.44**	**44.27**	**50.54**	85.31	92.05	**96.36**
GKL. MI	33.71	47.17	51.12	32.16	**44.87**	47.96	85.66	92.14	95.16

GC.MI was the fastest of the more complex algorithms

Influence of Sample Size

We use estimates

$$\hat{\Sigma}_N = \frac{1}{N} P^T P$$

For (sub)-Gaussian data we have[2]

$$\textit{If} \quad N \geq C(t/\epsilon)^2 d \qquad \textit{then} \quad \|\hat{\Sigma}_N - \Sigma\| \leq \epsilon$$

[2] with probability at least $1 - 2e^{-ct^2}$

Influence of Sample Size

We use estimates

$$\hat{\Sigma}_N = \frac{1}{N} P^T P$$

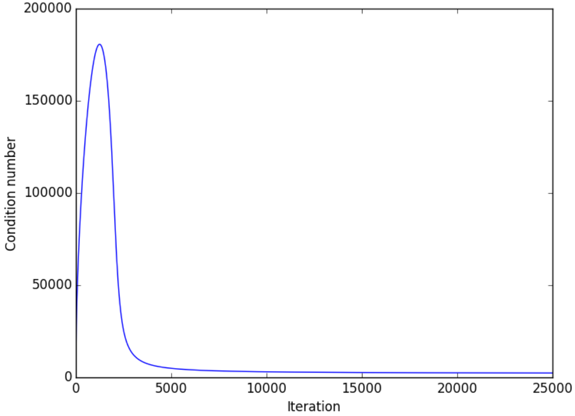
For (sub)-Gaussian data we have[2]

$$\textit{If} \quad N \geq C(t/\epsilon)^2 d \qquad \textit{then} \quad \|\hat{\Sigma}_N - \Sigma\| \leq \epsilon$$

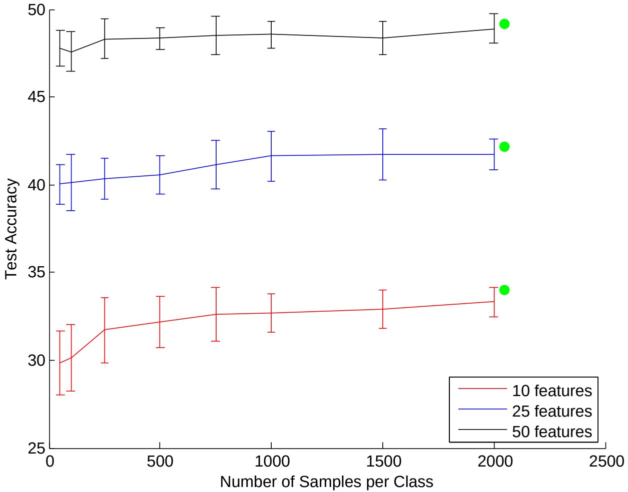
However the faster implementations use $\hat{\Sigma}_N^{-1}$

[2] with probability at least $1 - 2e^{-ct^2}$

Influence of Sample Size



$\kappa(\Sigma) = \frac{\frac{\|\Sigma^{-1}e\|}{\|\Sigma^{-1}b\|}}{\frac{\|e\|}{\|b\|}} = \|\Sigma^{-1}\|\|\Sigma\|$, for d $= 2048$ and various values of N



Effect of sample size on performance when using the Gaussian Approximation for the CIFAR dataset.

Jointly Informative Feature Selection Made Tractable by Gaussian Modeling

L.Lefakis and F.Fleuret

Journal of Machine Learning Research, 2016

The End

Thank You!