Feature Selection

Instructor: Hemanth Venkateswara

Computer Science & Engineering

ASU Arizona State University

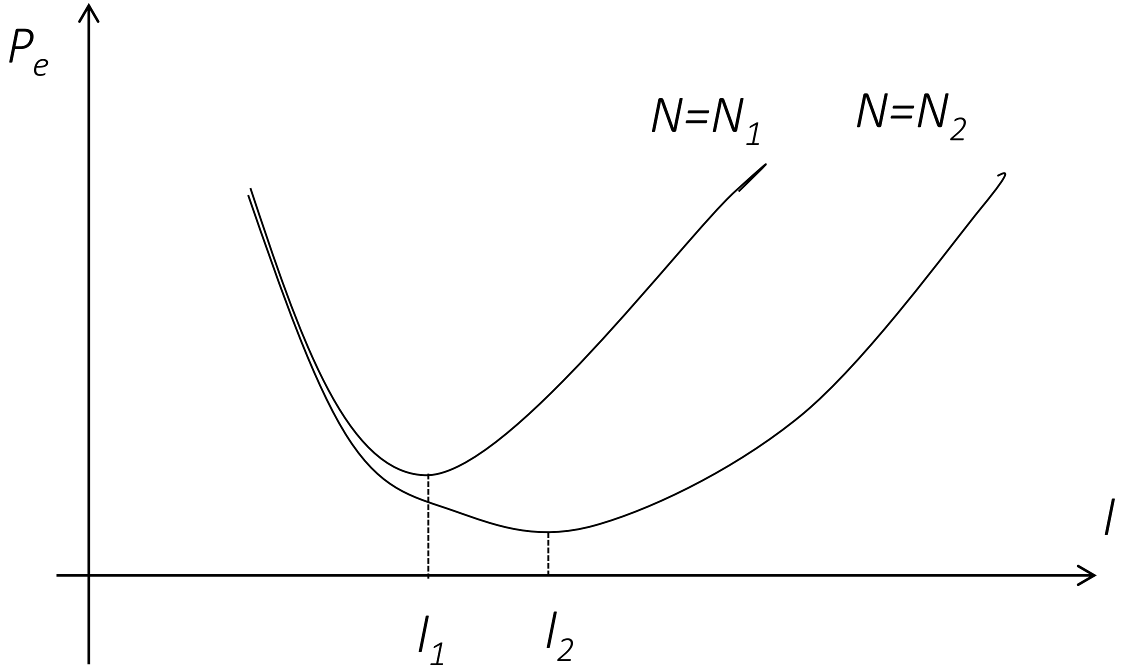
Overview

This topic discusses selecting an optimum number of features. We keep the best features (drop the rest) for improving classification accuracies

- Introduction
- Features selection based on hypothesis testing
- Scatter matrices
- L_1 regression (Lasso) feature selection
- Mutual information feature selection

Feature selection: Problem

- The goal: given a number of features, to select a "optimum" subset of l features (for, e.g., solving the same classification problem)
  - *The peaking phenomenon* ➔ there may exist an optimum l



N : size of training set

Basics of feature selection

- The goals:
  - Select the “optimum” number l of features
  - Select the “best” l features

- Large l has a three-fold disadvantage:
  - High computational demands
  - Low generalization performance
  - Poor error estimates

➔A caveat: there are classifiers whose performance is insensitive to the number l (or the dimensionality of the feature space)

Hypothesis testing

- The basic philosophy
  - Discard individual features with poor information content
  - The remaining information rich features are examined jointly as vectors

- Feature selection based on statistical Hypothesis Testing
  - The Goal: For each individual feature, find whether the values, which the feature takes for the different classes, differ significantly
    That is, answer
    - $H_1 : \theta \neq \theta_0$: The values differ significantly
    - $H_0 : \theta = \theta_0$: The values do not differ significantly

    If they do not differ significantly reject feature from subsequent stages.

Hypothesis testing basics

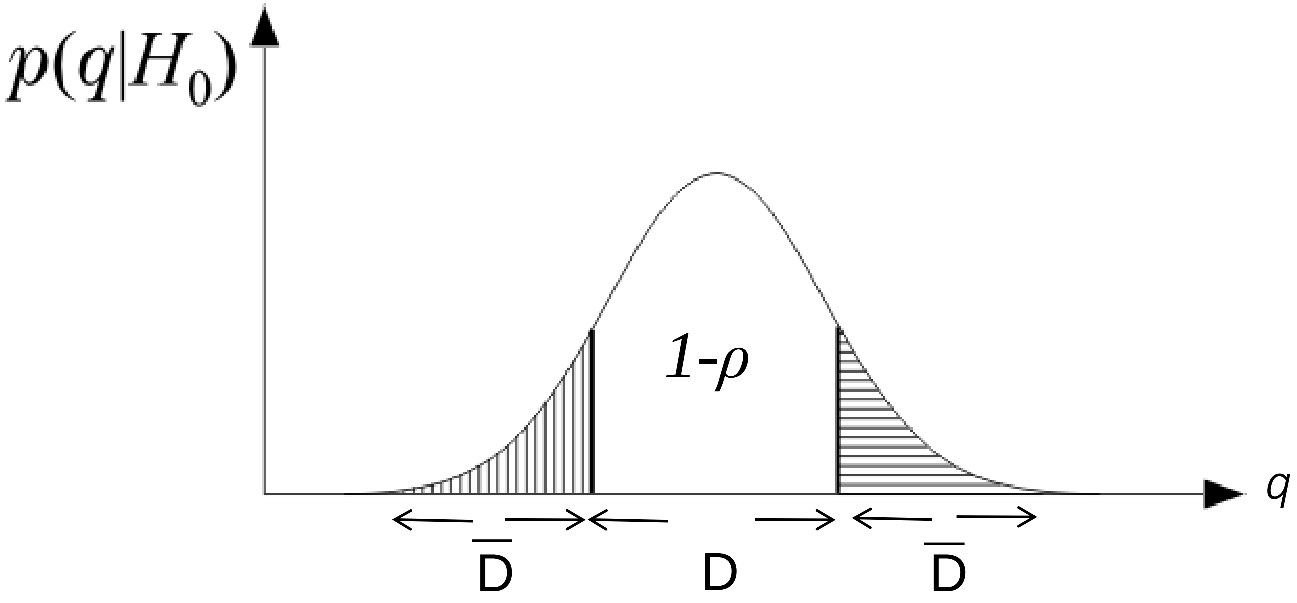
- The steps:
  - N measurements x_i for $i = 1,2,,\ldots,N$ are known
  - Define a function of them $q = f(x_1, x_2, \ldots, x_N)$: test statistic so that $p_q(q;\theta)$ is easily parameterized in terms of θ
  - Let D be an interval, where q has a high probability to lie under H_0 i.e., $p_q(q;\theta_0)$
  - Let $\overline{D}$ be the complement of D
    $D \rightarrow$ Acceptance Interval
    $\overline{D} \rightarrow$ Critical Interval

  - If q, resulting from $x_1, x_2, \ldots, x_N$ lies in D we accept H_0, otherwise we reject it

Hypothesis testing basics contd.

- Probability of error

$$p_q(q \in \bar{D} \big| H_0) = \rho$$



ρ is preselected and it is known as the significance level

An example

- Given $\{x_1, x_2, \dots, x_N\}$, assumed to be iid samples from certain pdf with ***unknown*** mean μ and ***known*** variance.
  - We want to test, for a given value $\mu_{0,}$

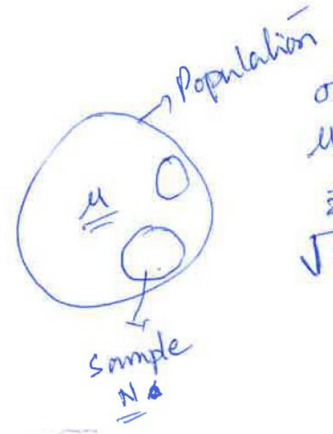
$$H_0: E[x] = \mu_0$$
$$H_1: E[x] \neq \mu_0$$

In words, if H_0 is true, it means we believe the samples are from a pdf with mean μ_0.

- A good q in this case will be $q = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$ where $\bar{x}$ is the sample mean.

Sampling distribution demo
http://onlinestatbook.com/stat_sim/sampling_dist/

Hypothesis testing basics contd.



$\sigma \to$ population std.

$\mu \to$ population mean.

$\bar{x} \to$ Sample mean.

$\sqrt{s^2} \to$ sample std.

$s^2 \to$ Sample variance.

Sampling distribution is the distribution of your measure over the samples.

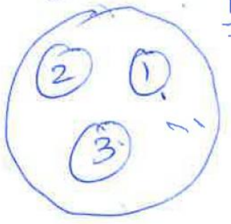
$\boxed{\mu_{\bar{x}}} \to$ sampling distribution mean

$\sigma_{\bar{x}} \to$ sampling distribution std.

sampling normal distribution

$S_1 \to \bar{x}_1 \to$ random variable

$S_2 \to \bar{x}_2$



$N = 2.$

	~~x~~ $\bar{x}$
① ②	→ 1.5
① ①	→ 1
② ③	→ 2.5
② ②	→ ②
② ①	→ 1.5
⋮	⋮

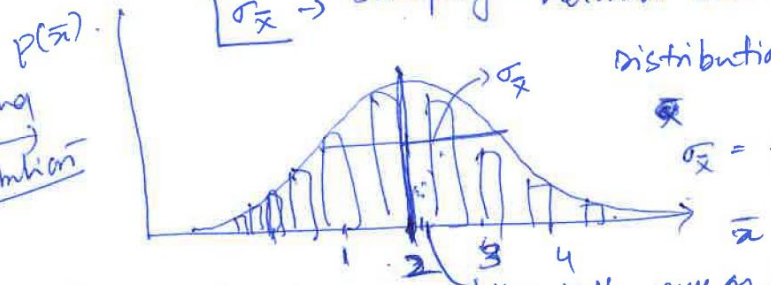
random variable

Sampling distribution

mean

max

median



Distribution over $\bar{x}$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

$\mu_{\bar{x}}$ is the same as μ.

Central Limit theorm

$E[\bar{x}] = \mu.$ population mean.

Sampling distribution standard deviation $\sqrt{\sigma_{\bar{x}}^2}$

~~pon~~

Sampling distribution example – hypothesis testing

- A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response time is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think the drug has an effect on response time?

H_0 : Drug has no effect.

H_1 : Drug has effect. → $\mu \neq 1.2$ sec with drug.

$N = 100$. $\mu_{\bar{x}} = \bar{x} = 1.05$ $\quad \mu = 1.2$

σ = population std.

Sampling distribution →

std. of the sampling distribution $= \sigma_{\bar{x}}$

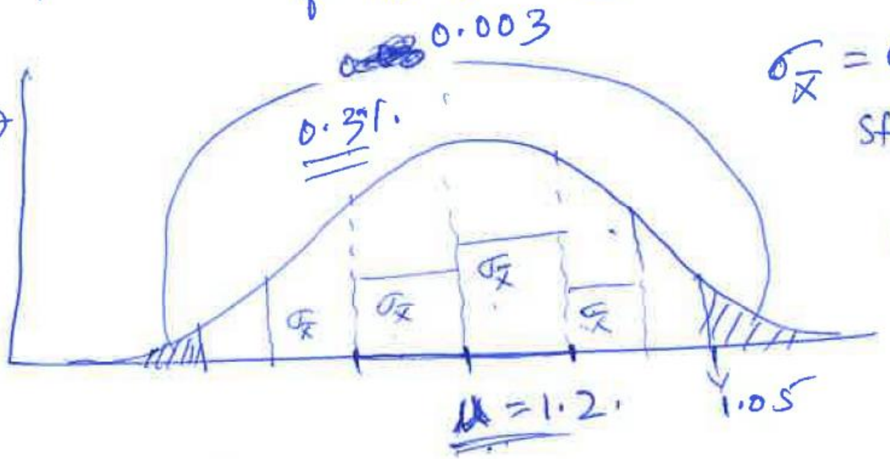
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \approx \frac{\sqrt{s^2}}{\sqrt{N}} = \frac{0.5}{\sqrt{100}} = 0.05$$

$\sqrt{s^2}$ → sample std $= 0.5$ sec.

if $N \geq 30$ we can set $\sqrt{s^2} \approx \sigma$

Sample std. $\sqrt{s^2}$ = population std σ.

What is the probability of getting $\mu_{\bar{x}} = 1.05$ if H_0 is true.



$\sigma_{\bar{x}} = 0.05$ is the std. of the sampling distribution.

$$q = \frac{\mu_* - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$\frac{1.2 - 1.05}{0.05} \approx 3$$

with 1 std $\approx$ 68%.

with 2 std $\approx$ 95%.

with 3 std = 99.7%.

$\mu_{\bar{x}} = 1.05$

$\mu_{\bar{x}} = 1.05$ lies 3 std. away from μ

Problem credit: Khan academy

Application to feature selection: An example

- E.g., to test against zero the difference $\mu_1 - \mu_2$ of the respective means in ω_1, ω_2 of a single feature.
  - Let $x_1, x_2, \ldots, x_N$, the values of a feature in ω_1
  - Let $y_1, y_2, \ldots, y_N$, the values of the same feature in ω_2
  - Assume in both classes $\sigma_1^2 = \sigma_2^2 = \sigma^2$

- The test becomes

$$H_0: \Delta\mu = \mu_1 - \mu_2 = 0$$
$$H_1: \Delta\mu \neq 0$$

Feature selection example

- Example: The values of a feature in two classes are:

ω_1: 3.5, 3.7, 3.9, 4.1, 3.4, 3.5, 4.1, 3.8, 3.6, 3.7
ω_2: 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.1, 3.3, 3.6

- The significance level is ρ=0.05
- We have $\omega_1: \; \bar{x} = 3.73, \; \hat{\sigma}_1^2 = 0.0601$

$\omega_2: \; \bar{y} = 3.25, \; \hat{\sigma}_2^2 = 0.0672$

$$H_0 : \Delta\hat{\mu} = 0$$
$$H_1 : \Delta\hat{\mu} \neq 0$$

$$S_z^2 = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$$

$$q = \frac{(\bar{x} - \bar{y}) - 0}{S_z \sqrt{\frac{2}{10}}}$$

$$q = 4.25$$

From the table of the t-distribution with 2*N*-2=18 degrees of freedom and ρ=0.05, we obtain *D*=[-2.10,2.10] and since *q*=4.25 is outside *D*, H_1 is accepted and the feature is selected.

Some remarks

- The examples we considered thus far are for treating the features independently: In general, feature may be correlated

- If variance is known, we use a standard z-score. But, with unknown variances, we may end up with dealing with t-distribution, $\mathcal{X}^2$ distribution, or F distribution etc. But the general idea of hypothesis testing is still useful.

- The difference of mean values may be useful for helping decide on discarding features, only when the spread around the mean is not too big as to blur the class distinction
  - In general, we may further consider the ROC curve: a good feature should result in an ROC curve further away from the straight line.

Class separability measures

- Two features may be rich in information, but if they are highly correlated we need not consider both of them.

- A more general approach:
  - Choose the number, *l*, of features to be used. This is dictated by the specific problem (e.g., the number, *N*, of available training patterns and the type of the classifier to be adopted).
  - Combine remaining features to search for the “best” combination.
  - Two possible approaches to this:
    1. Use different feature combinations to form the feature vector. Train the classifier, and choose the combination resulting in the best classifier performance.
    2. Adopt a class separability measure and choose the best feature combination against this cost.

Divergence

- Let $\underline{x}$ be the current feature combination.
- Consider the two–class case. Obviously, if on average the value of $\ln \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)}$ is close to zero, then $\underline{x}$ is a poor feature combination

- Define

$$D_{12} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_1) \ln \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} d\underline{x}$$

$$D_{21} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_2) \ln \frac{p(\underline{x}|\omega_2)}{p(\underline{x}|\omega_1)} d\underline{x}$$

$$d_{12} = D_{12} + D_{21}$$

- d_{12} is known as the divergence and can be used as a class separability measure.

Scatter matrices

- Within-class scatter matrix S_w, between-class scatter matrix S_B , and mixture scatter matrix $S_m = S_w + S_B$.

- Measures based on Scatter Matrices:

$$J_1 = \frac{\text{Trace}\{S_m\}}{\text{Trace}\{S_w\}} \qquad J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$$

$$J_3 = \text{Trace}\{S_w^{-1} S_m\}$$

- The above J_1, J_2, and J_3 take high values for the cases where: data are clustered together within each class and/or the means of the various classes are far.

Fisher’s discriminant ratio

- Fisher’s discriminant ratio. In one dimension and for two equiprobable classes the determinants become:

$$|S_w| \propto \sigma_1^2 + \sigma_2^2$$

$$|S_b| \propto (\mu_1 - \mu_2)^2$$

and

$$\frac{|S_b|}{|S_w|} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

known as Fisher’s ratio.

➔ can be extended to multiple classes.

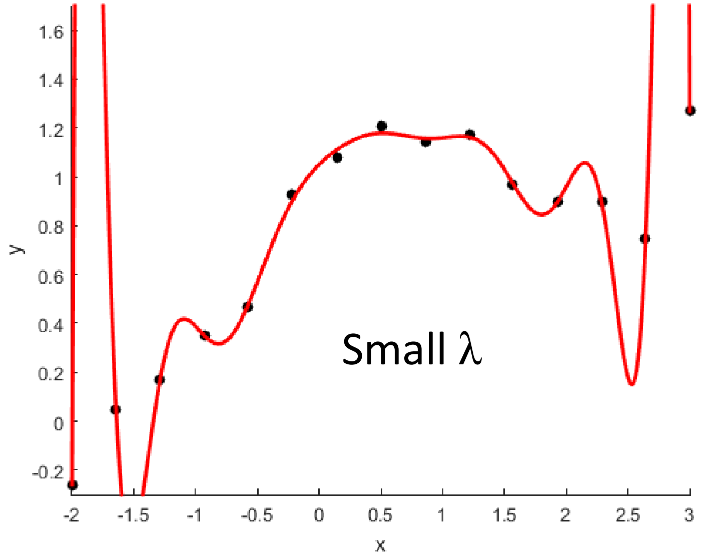
Ridge regression with polynomials

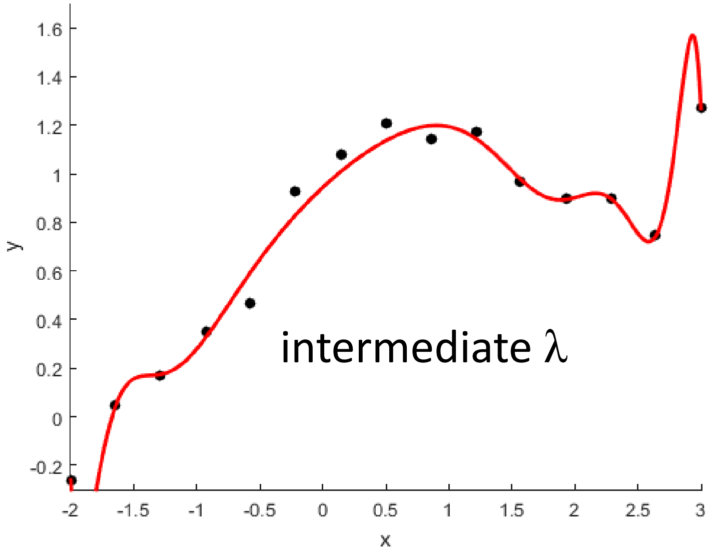
Consider fitting a polynomial of degree 15 to the data. Let $\boldsymbol{x} \in \mathbb{R}$, then

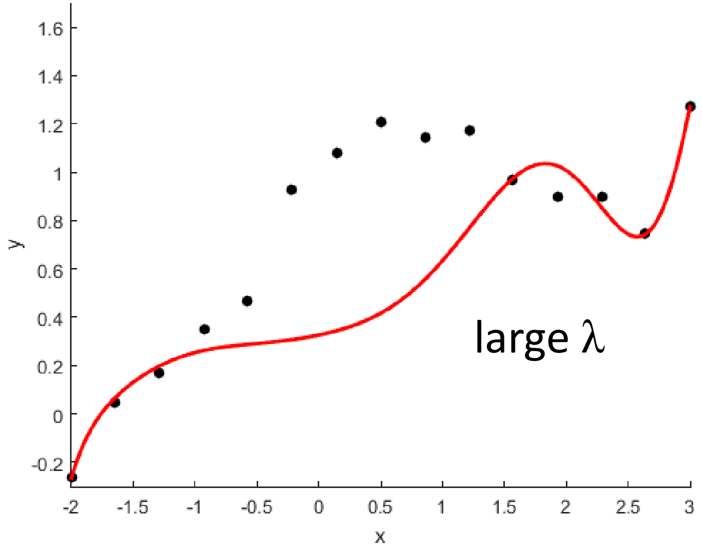
$$\hat{y} = \theta_0 + \theta_1 \boldsymbol{x} + \theta_2 \boldsymbol{x}^2 + \ldots + \theta_{14} \boldsymbol{x}^{14} + \theta_{15} \boldsymbol{x}^{15}$$

$$\phi(\boldsymbol{x}) = [1, \boldsymbol{x}, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{14}, \boldsymbol{x}^{15}]^\top$$

Loss function $$J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{N} \{y_i - \phi(x_i)^T \boldsymbol{\theta}\}^2 + \frac{\lambda}{2} \boldsymbol{\theta}^T \boldsymbol{\theta}$$





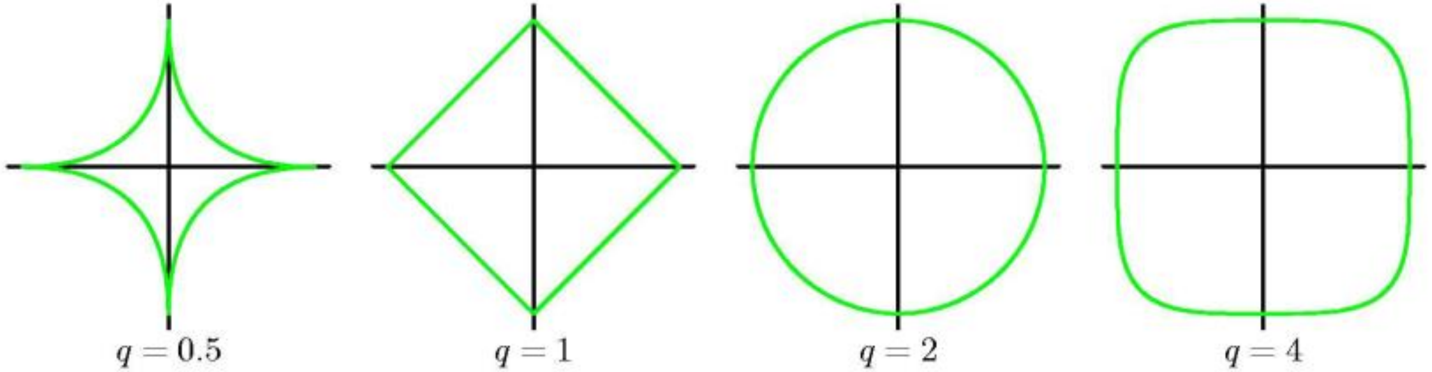


General regularizer

- A more general regularizer is:

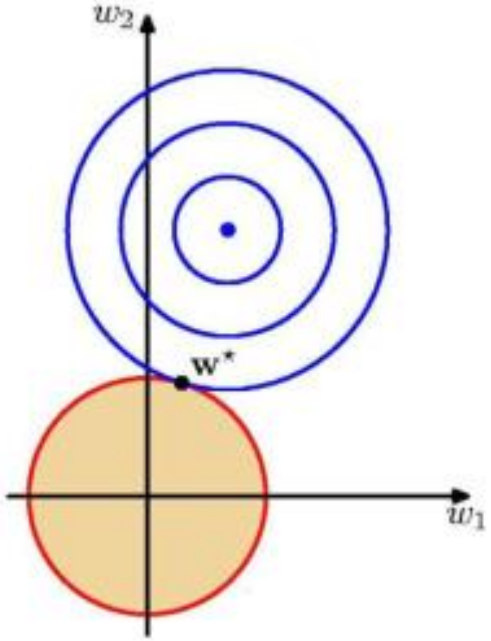
$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N}\{y_i - \phi(x_i)^T\boldsymbol{\theta}\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|\theta_j|^q$$

- Ridge regression q=2
- L_1 norm - Lasso: q=1

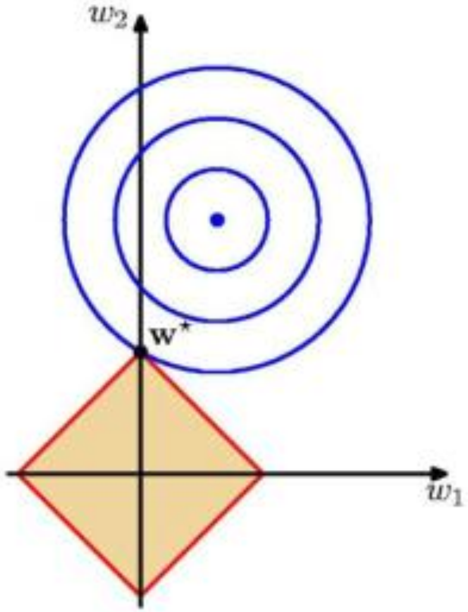


Why Lasso induces sparsity

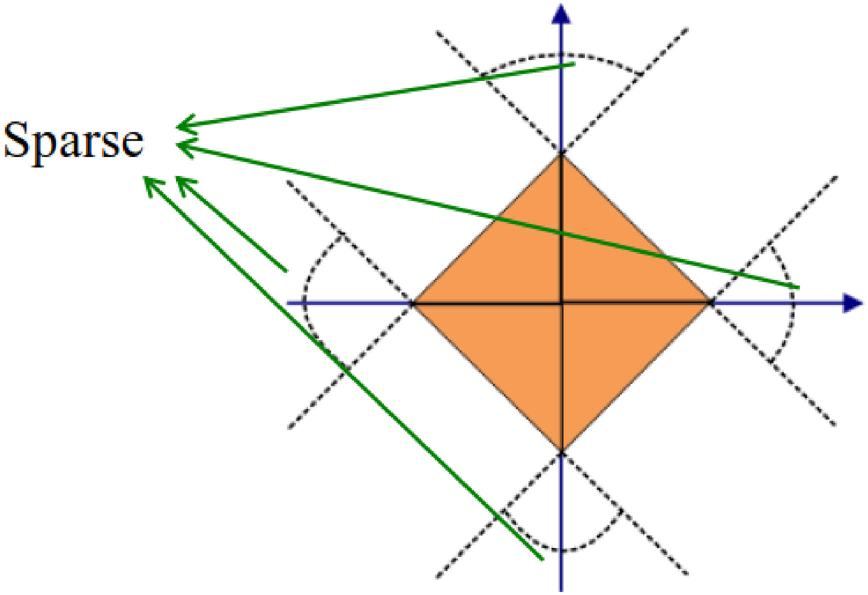
- Probability of intersection of loss function with the Lasso constraint is higher at the vertices



Ridge or L_2 regression



Lasso or L_1 regression



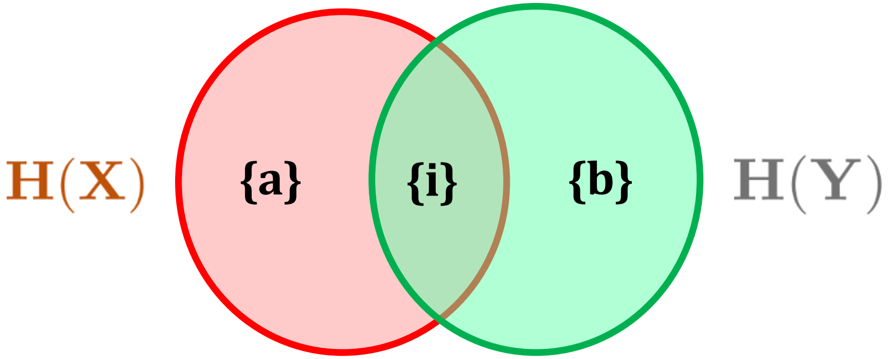
At the vertices, components of $\boldsymbol{\theta}$ (or $\boldsymbol{w}$ in the figure) are zero. The features of $\phi(x)$ corresponding to those $\boldsymbol{\theta}$ components are therefore dropped. The Lasso regression performs implicit feature selection.

Mutual information

Information $I(X = \mathbf{x}) = -\log p(X = \mathbf{x})$

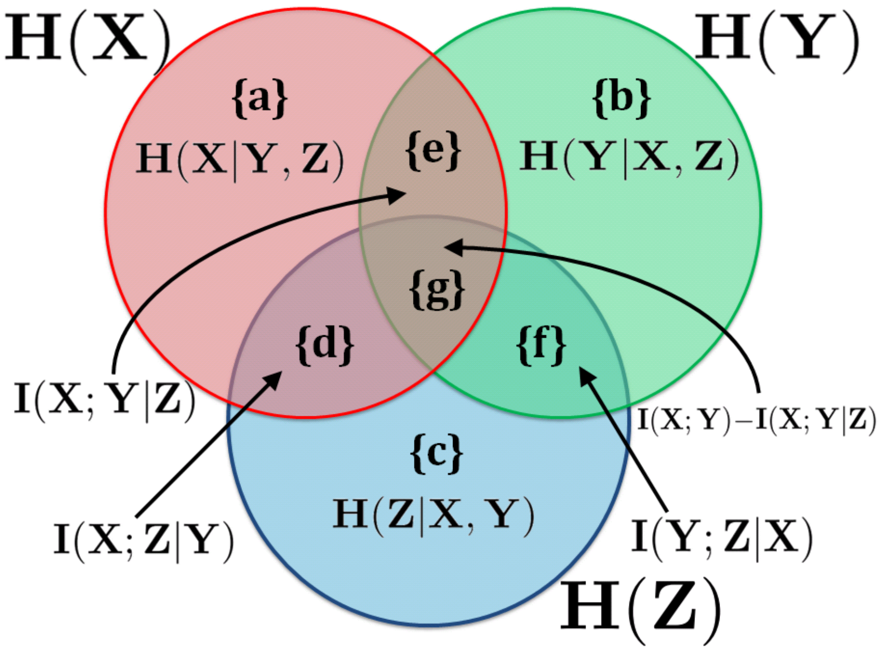
Expected information $H(X) = \mathbb{E}[I(X)]$

Entropy $H(X) = -\sum_{\mathbf{x}} p(X = \mathbf{x}) \log p(X = \mathbf{x})$



{a} = $H(X|Y)$ {b} = $H(Y|X)$

{i} = $I(X;Y)$ Mutual Information (MI)



$I(X;Y|Z)$ Conditional Mutual Information (CMI)

H(X)= {a,e,g,d}, H(Y)={b,e,g,f}, H(Z)={c,d,g,f},

I(X;Y)={e,g}, I(X;Z) = {d,g}, I(Y;Z) = {g,f},

H(X,Y,Z) = {a,b,c,d,e,f,g}

Mutual information based feature selection

$$\mathscr{D} = \{(\mathbf{x}^i, y^i); i = 1\dots m\}:\ \mathbf{x}^i \in \mathbb{R}^n,\ y^i \in \{1, \dots, c\}$$

$\mathbf{x}^i$, is an instance of n continuous random variables

$$X = \{X_1, X_2, \dots, X_n\}$$

y^i, is an instance of discrete random variable Y

$\mathbb{S}$ is subset of k feature indices, $1 \leq k \leq n$

$$X_{\mathbb{S}} \subseteq X, \quad X_{\tilde{\mathbb{S}}} = X \backslash X_{\mathbb{S}}$$

Theorem: $p(Y|X) = p(Y|X_{\mathbb{S}})$, iff $I(X;Y) = I(X_{\mathbb{S}};Y)$

Assumption to Resolve Intractable $I(X_{\mathbb{S}};Y)$

Let $\mathbb{S}_i = \mathbb{S} \backslash \{i\}$

Assumption:

$$p(X_{\mathbb{S}_i}|X_i) = \prod_{j \in \mathbb{S}_i} p(X_j|X_i)$$

$$p(X_{\mathbb{S}_i}|X_i, Y) = \prod_{j \in \mathbb{S}_i} p(X_j|X_i, Y)$$

Conditional mutual information

Theorem: $I(X_{\mathbb{S}};Y) = I(X_i;Y) + \sum_{j \in \mathbb{S}_i} I(X_j;Y|X_i)$

For details refer to: Efficient Approximate Solutions to Mutual Information Based Global Feature Selection, Venkateswara et al. ICDM 2016

Conditional mutual information feature selection

$$\mathbb{S} = \underset{\{\mathbb{S}|X_{\mathbb{S}}\subset X\},|\mathbb{S}|=k}{\operatorname{argmax}} \sum_{i\in\mathbb{S}} \left[I(X_i;Y) + \sum_{j\in\mathbb{S}_i} I(X_j;Y|X_i) \right]$$

Global Solution

x is binary vector
1 indicates
selected feature

x_1	0
:	:
:	:
x_i	1
:	:
:	:
x_j	1
:	:
x_d	0

This is equivalent to the constrained Binary Quadratic problem,

$$\max_{\mathbf{x}}\{\mathbf{x}^\top \mathbf{Q}\mathbf{x}\} \quad \text{s.t. } \mathbf{x} \in \{0,1\}^n, \ \ ||\mathbf{x}||_1 = k, \qquad \text{(BQP)}$$

NP hard problem

$\mathbf{Q}$ is a $[n \times n]$ non-negative matrix , $Q_{ii} = I(X_i;Y)$, $Q_{ij} = I(X_j;Y|X_i)$
$\mathbb{S} =$ non-zero indices of the solution $\mathbf{x}$.

There are techniques to solve this non-convex problem such as Spectral programming, Semi-definite programming, Truncated Power Method and Low-Rank Bilinear Approximation

For details refer to: Efficient Approximate Solutions to Mutual Information Based Global Feature Selection, Venkateswara et al. ICDM 2016