Lecture 16

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Continuing from last lecture, where we wanted to find a transformation function so that it converts to an easier version.

> which is equivalent to,
>
> $$q(\hat{x}) = p(x)\left|\frac{dx}{d\hat{x}}\right|$$

Now, moving on

Feature Reduction:

- **Filter**: For each feature decide keep v/s discard. Then train your model with the kept features $\implies$ Best Subset Selection over multiple subsets
- **Wrapper**: For some n subsets, Forward Selection and Backward Elimination

Forward Selection:
start with an empty set S, set of remaining features R include all x_i's (i goes from 1 to D) at this point.

loop over j in D: `Note: alternative stopping criteria are allowed`
loop over elements x_i of R:
evaluate utility of including x_i in S
pick the best feature, include in S, remove from R
you get the subset features S and R

Backward Elimination:
S is the whole set and R is empty

loop over j:
loop over elements x_k of S:
evaluate utility of removing x_j from S
remove the worst feature from S, add to R
you get the subset S and R

- **Embedded**:
  Examples: LASSO (L1), Elastic Net (L1+L2)
  Learning and Subset Selection are integrated.

L2 SVM:

$$\frac{\lambda}{2}\|w\|_2^2 + \sum_i [1 - t_i y_i]_+ \text{ Hinge Loss}$$

L1 SVM:

$$\lambda\|w\|_1 + \sum_i [1 - t_i y_i]_+ \text{ Hinge Loss}$$

$$= \lambda \sum_j |w_j| + \sum_i [1 - t_i(w^T x_i + b)]_+ \text{ Hinge Loss}$$

where the first term represents sparsity in j and the other term represents sparsity in i

Principal Component Analysis

`Feature Reduction Technique`
$x_3 = ax_1 + bx_2$

Now, starting with Kernelized SVMs

Kernelized Support Vector Machines

`Detour:`
$y = f(w_1x_1 + w_2x_2 + \ldots + w_Dx_D + w_0)$
If Gaussian distribution is assumed, then perform standard normalization (mean = 0, variance = 1).
If you assume uniform, perform min-max normalization.

Objectives:

- dual form of SVM
- replacing raw data with kernels
- derive support vector regression objective

Kernel - way to compare two samples in similarity.
$k(x, \hat{x}) \rightarrow$ high

$k(x, \tilde{x}) \rightarrow$ low

Example:

1. $k(x, \hat{x}) = exp(-a\|x - \hat{x}\|_2^2)$: Gaussian Kernel (Radial Basis Function (RBF))
2. $\frac{|x.\hat{x}|}{\|x\|\|\hat{x}\|}$ Cosine Kernel