Feature selection

Usman Roshan

What is feature selection?

- Consider our training data as a matrix where each row is a vector and each column is a dimension.
- For example consider the matrix for the data x_1=(1, 10, 2), x_2=(2, 8, 0), and x_3=(1, 9, 1)
- We call each dimension a feature or a column in our matrix.

Feature selection

- Useful for high dimensional data such as genomic DNA and text documents.
- Methods
  - Univariate (looks at each feature independently of others)
    - Pearson correlation coefficient
    - F-score
    - Chi-square
    - Signal to noise ratio
    - And more such as mutual information, relief
  - Multivariate (considers all features simultaneously)
    - Dimensionality reduction algorithms
    - Linear classifiers such as support vector machine
    - Recursive feature elimination

Feature selection

- Methods are used to rank features by importance
- Ranking cut-off is determined by user
- Univariate methods measure some type of correlation between two random variables. We apply them to machine learning by setting one variable to be the label (y_i) and the other to be a fixed feature (x_{ij} for fixed j)

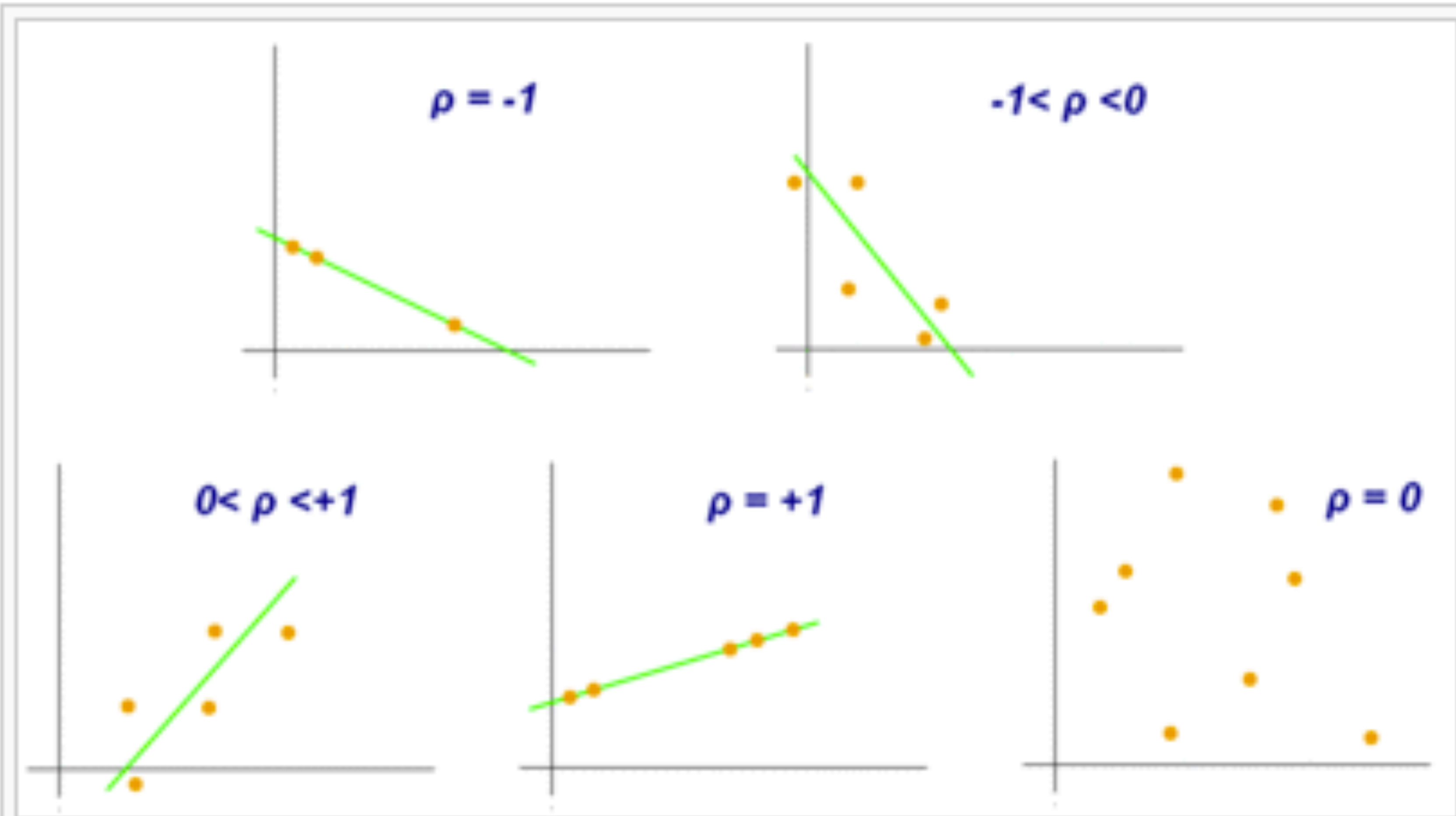
Pearson correlation coefficient

- Measures the correlation between two variables
- Formulas:
  - Covariance(X,Y) = E((X-μ_X)(Y-μ_Y))
  - Correlation(X,Y)= Covariance(X,Y)/$\sigma_X\sigma_Y$
  - Pearson correlation =

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- The correlation r is between -1 and 1. A value of 1 means perfect positive correlation and -1 in the other direction

Pearson correlation coefficient



From Wikipedia

F-score

F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $\boldsymbol{x}_k, k = 1, \ldots, m$, if the number of positive and negative instances are n_+ and n_-, respectively, then the F-score of the ith feature is defined as:

$$F(i) \equiv \frac{\left(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i\right)^2 + \left(\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i\right)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)}\right)^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)}\right)^2}, \qquad (4)$$

where $\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{x}}_i^{(+)}$, $\bar{\boldsymbol{x}}_i^{(-)}$ are the average of the ith feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. Therefore, we use this score as a feature selection criterion.

From Lin and Chen, Feature extraction, 2006

Chi-square test

- We have two random variables:
  - Label (L): 0 or 1
  - Feature (F): Categorical
- Null hypothesis: the two variables are independent of each other (unrelated)
- Under independence
  - P(L,F)= P(D)P(G)
  - P(L=0) = (c1+c2)/n
  - P(F=A) = (c1+c3)/n
- Expected values
  - E(X1) = P(L=0)P(F=A)n
- We can calculate the chi-square statistic for a given feature and the probability that it is independent of the label (using the p-value).
- Features with very small probabilities deviate significantly from the independence assumption and therefore considered important.

Contingency table

	Feature=A	Feature=B
Label=0	Observed=c1 Expected=X1	Observed=c2 Expected=X2
Label=1	Observed=c3 Expected=X3	Observed=c4 Expected=X4

Signal to noise ratio

- Difference in means divided by difference in standard deviation between the two classes
- $S2N(X,Y) = (\mu_X - \mu_Y)/(\sigma_X + \sigma_Y)$
- Large values indicate a strong correlation

Multivariate feature selection

- Consider the vector w for any linear classifier.
- Classification of a point x is given by wTx+w0.
- Small entries of w will have little effect on the dot product and therefore those features are less relevant.
- For example if w = (10, .01, -9) then features 0 and 2 are contributing more to the dot product than feature 1. A ranking of features given by this w is 0, 2, 1.

Multivariate feature selection

- The w can be obtained by any of linear classifiers we have see in class so far
- A variant of this approach is called recursive feature elimination:
  - Compute w on all features
  - Remove feature with smallest w_i
  - Recompute w on reduced data
  - If stopping criterion not met then go to step 2

Feature selection in practice

- NIPS 2003 feature selection contest
  - Contest results
  - Reproduced results with feature selection plus SVM
- Effect of feature selection on SVM
- Comprehensive gene selection study comparing feature selection methods
- Ranking genomic causal variants with SVM and chi-square

Limitations

- Unclear how to tell in advance if feature selection will work
  - Only known way is to check but for very high dimensional data (at least half a million features) it helps most of the time
- How many features to select?
  - Perform cross-validation