CS 145 Discussion 1

Linear Regression

Reminders

- 10/11/2017 (Next Wednesday)
  - HW1 out, due 10/18/2017 (1 week)
  - Group formations for course project due
- New TA hours schedule
  - 12:30-2:30 AM on Tuesdays
  - 9:30-11:30 AM on Thursdays
  - 2:30-4:30 PM on Thursdays
  - @ BH 2432

Overview

- Feature extraction from real data
- Linear regression application
- Matrix form of least square estimation
- Math review

Feature Extraction from Real Data

- Types of Features
  - Numerical
  - Categorical
    - Nominal, Binary, Ordinal

- Real data may be messy for extracting features
  - Unorganized structure
  - Hidden and deep information


```
{
  "created_at": "Tue Nov 24 00:14:03 +0000 2015",
  "id": 668945640508911600,
  "id_str": "668945640508911616",
  "text": "I know that I let you down. Is it too late now to say sorry?",
  "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 285176507,
    "id_str": "285176507",
    "name": "NINA XELYN",
    "screen_name": "nxelyn",
    "location": "Norway/CA",
    "url": "http://instagram.com/ninaxelyn",
    "description": "sci nxelyn | gemini | sjsu",
    "protected": false,
    "verified": false,
    "followers_count": 232,
    "friends_count": 64,
    "listed_count": 1,
    "favourites_count": 842,
    "statuses_count": 1332,
    "created_at": "Wed Apr 20 17:28:57 +0000 2011",
    "utc_offset": -28800,
    "time_zone": "Pacific Time (US & Canada)",
    "geo_enabled": true,
    "lang": "en",
    "contributors_enabled": false,
    "is_translator": false,
    "profile_background_color": "C6E2EE",
    "profile_background_image_url": "http://abs.twimg.com/images/themes/theme2/bg.gif",
    "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme2/bg.gif",
    "profile_background_tile": false,
    "profile_link_color": "1F98C7",
    "profile_sidebar_border_color": "C6E2EE",
    "profile_sidebar_fill_color": "DAECF4",
    "profile_text_color": "663B12",
    "profile_use_background_image": true,
    "profile_image_url": "http://pbs.twimg.com/profile_images/668763072245268480/qFwM2HjH_normal.jpg",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/668763072245268480/qFwM2HjH_normal.jpg",
    "profile_banner_url": "https://pbs.twimg.com/profile_banners/285176507/1444614000",
    "default_profile": false,
    "default_profile_image": false,
    "following": null,
    "follow_request_sent": null,
    "notifications": null
  },
  "geo": null,
  "coordinates": null,
  "place": {
    "id": "7d62cffe6f98f349",
    "url": "https://api.twitter.com/1.1/geo/id/7d62cffe6f98f349.json",
    "place_type": "city",
    "name": "San Jose",
    "full_name": "San Jose, CA",
    "country_code": "US",
    "country": "United States",
    "bounding_box": {
      "type": "Polygon",
      "coordinates": [
        [
          [
            -122.035311,
            37.193164
          ],
          [
            -122.035311,
            37.469154
          ],
          [
            -121.71215,
            37.469154
          ],
          [
            -121.71215,
            37.193164
          ]
        ]
      ]
    },
    "attributes": {}
  },
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 0,
  "favorite_count": 0,
  "entities": {
    "hashtags": [],
    "urls": [],
    "user_mentions": [],
    "symbols": []
  },
  "favorited": false,
  "retweeted": false,
  "filter_level": "low",
  "lang": "en",
  "timestamp_ms": "1448324043763"
}
```


Example of a tweet data

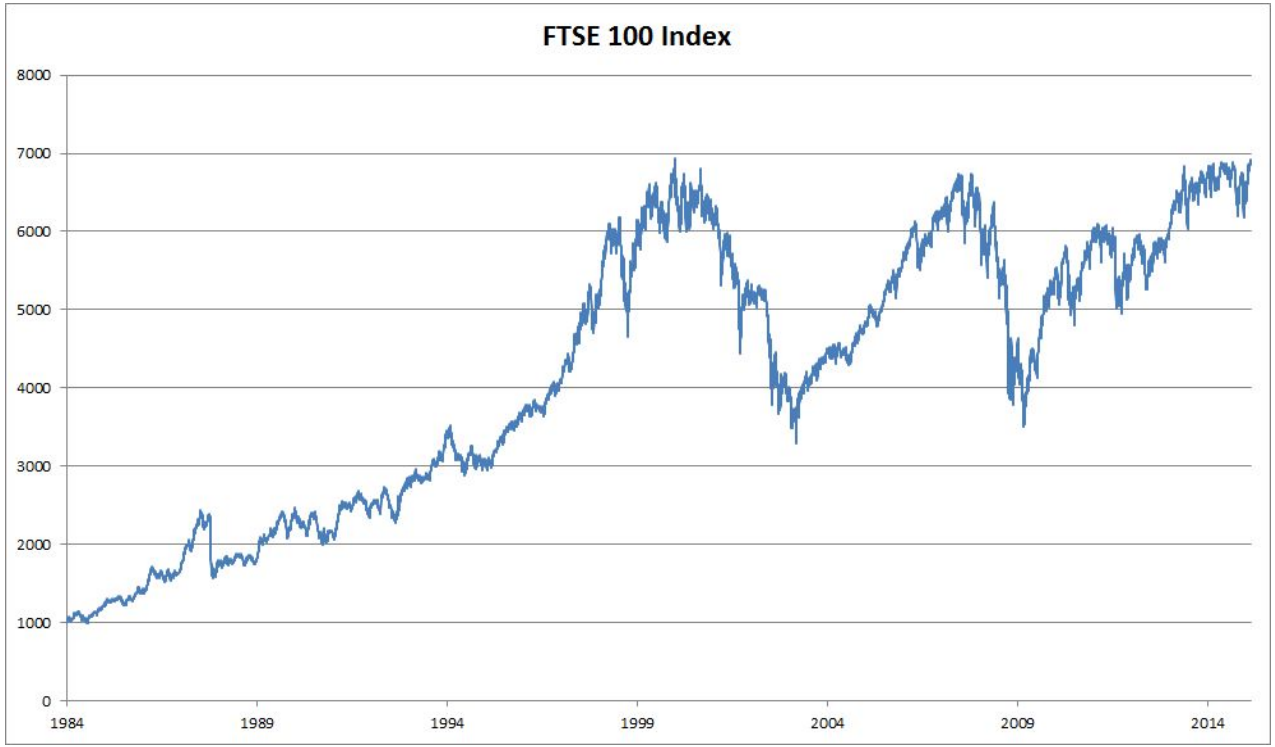
Numerical Features

- Numerical attributes
  - Raw data with numerical formats
  - E.g., numbers of friends and followers, timestamps
- Numerical statistics
  - Numerical statistics towards a characteristic
  - E.g., the length of text, the average daily number of tweets for the user
- Numerical hidden representations
  - Represent data in optimized hidden spaces
  - E.g, pLSA and LDA for text (Week 10)

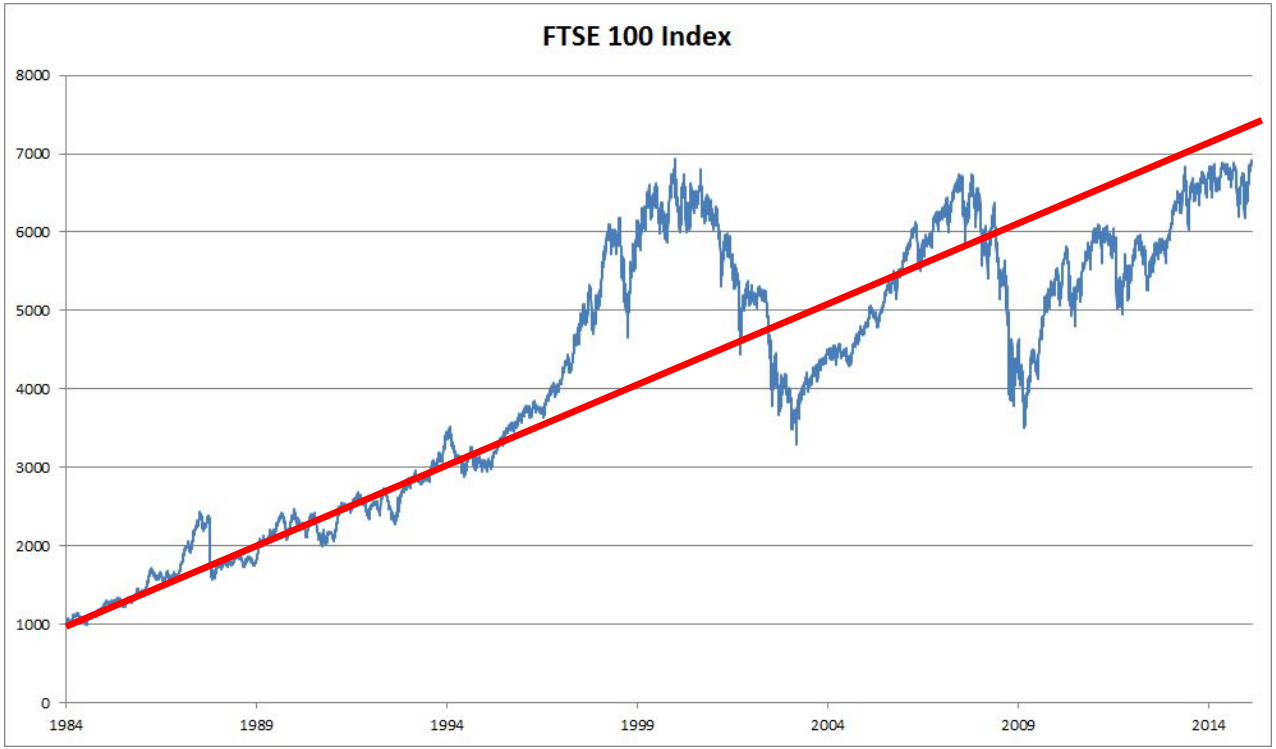
Categorical Features

- Categorical attributes
  - Raw data which originally have a set of discrete categories
  - E.g., cities of users, languages of text,
- Discretization for numerical attributes
  - Transform numerical features into categorical features
  - E.g., Morning/Afternoon/Night, Long/Short Text (more than k words?)
- Categorical statistics
  - Categorical statistics towards a characteristic
  - E.g., If the user posts more than k tweets in a week, Few/Usual/Many tweets posted in near regions

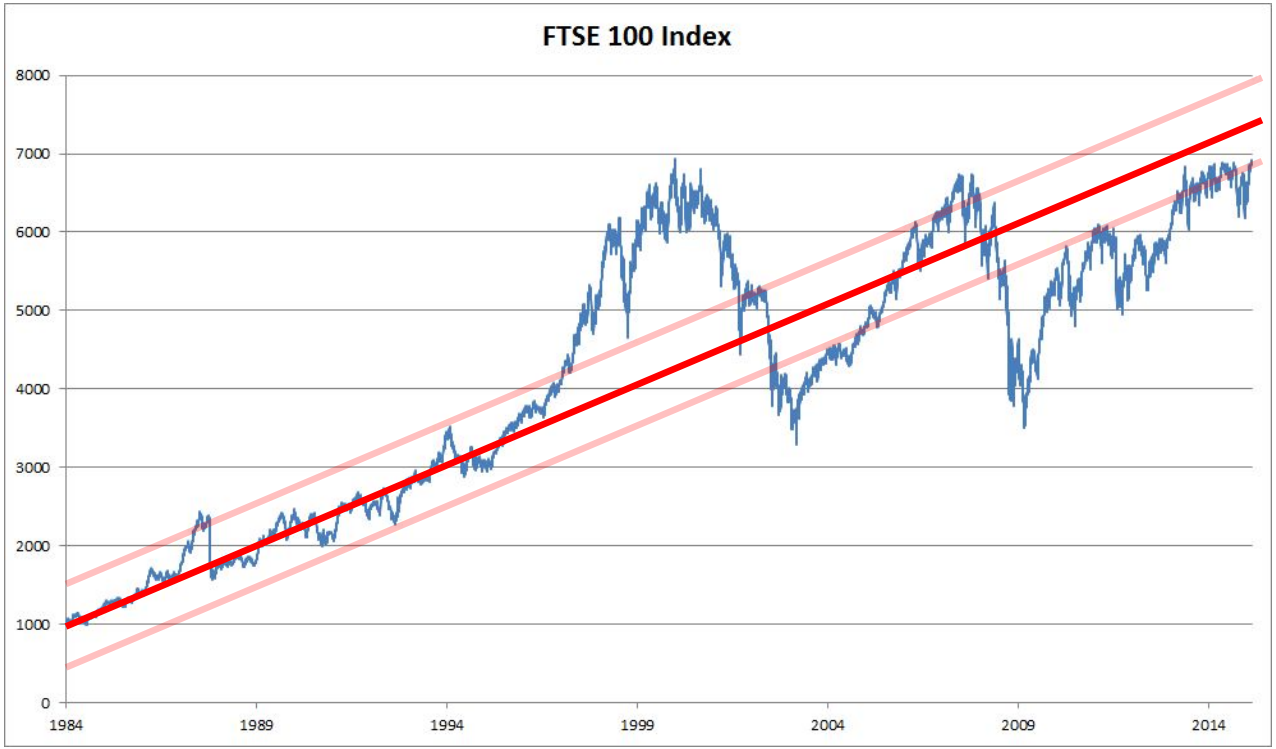
An application of linear regression: Stock Prices



An application of linear regression: Stock Prices

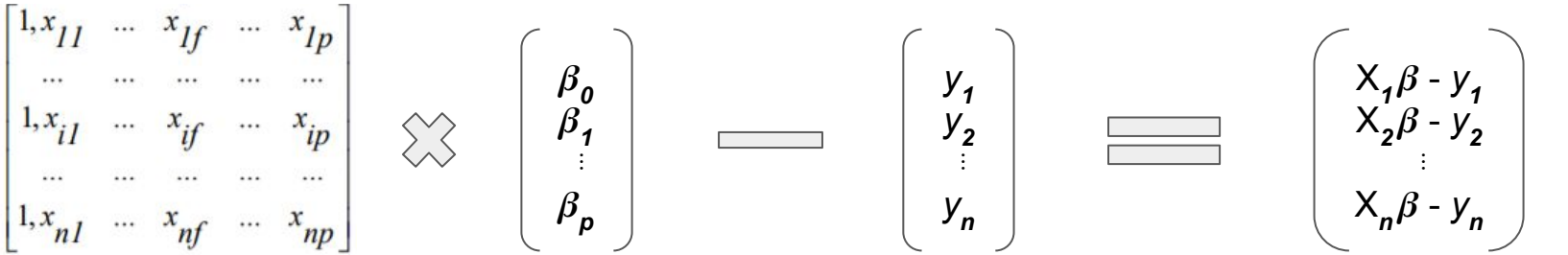


An application of linear regression: Stock Prices



Matrix form of Least Square Estimation

$$J(\boldsymbol{\beta}) = (\mathrm{X}\boldsymbol{\beta} - \boldsymbol{y})^T (X\boldsymbol{\beta} - \boldsymbol{y})/2$$



$\boldsymbol{X}$: $\boldsymbol{n} \times (\boldsymbol{p}+\boldsymbol{1})$ matrix

β: (p + 1) x 1 matrix

y: n x 1 matrix

Xβ - y: n x 1 matrix

$$J(\boldsymbol{\beta}) = \left|\left|\mathrm{X}\boldsymbol{\beta} - \boldsymbol{y}\right|\right|^2/2$$