Support Vector Machines: we want to maximize:

$$\text{Margin} = \frac{2}{||w||^2}$$

Constraints:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

If the problem is not linearly separable, we should introduce slack variables:
We need to minimize:

Minimize:

$$L(w) = \frac{||\vec{w}||^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k\right)$$

It is Subject to:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 - \xi_i \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 + \xi_i \end{cases}$$

Boosting: An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records: Initially, all N records are assigned equal weights and Unlike bagging, weights may change at the end of boosting round.

Recommender Systems

Neighborhood Methods:
Pros:
- Intuitive / easy to explain
- No training
- Handles new users/items

Challenges:
- Users rate differently (bias)
- Ratings change over time (bias)

Feature Extraction - Content-Base:
Realistically:
- It's difficult to characterize movies and users with the right features
- Characterization of users and movies may not be accurate

Feature Extraction - Collaborative Filtering:
Can't use SVD because R is sparse… BUT, we can formulate an optimization problem to solve:

$$\min_{p,q} \sum_{i,j \in R} (r_{ij} - p_i^T q_j)^2 + \lambda(\|p\|_F^2 \|p\|_F^2)$$

To solve, take derivatives wrt P & Q. Then, just like Expectation-Maximization Algorithm from GMM:
1. Start with random Q
2. Get P
3. Improve Q
4. Repeat 2 & 3

Linear Regression

Challenge for those who have LR experience:
○ Every day my alarm goes off at seemingly random times…
○ I've recorded the times for the past year of so (1 - 355 days)
○ Today is day 356
○ Can you predict when my alarm will ring?

Motivation:
Suppose we are given a curve y = h(x), how can we evaluate whether it is a good fit to our data?
Compare h(xi) to yi for all i.
Goal: For a given distance function d, find h where L is smallest.

$$L(h) = \sum_i d(h(x_i), y_i)$$

Another way to define this problem is in terms of probability.
Define P(Y | h) as the probability of observing Y given that it was sampled from h.
Goal: Find h that maximizes the probability of having observed our data.

Minimize:

$$L(h) = \sum_i d(h(x_i), y_i)$$

Maximize:

$$L(h) = P(Y \mid h)$$

Let's start by assuming our data was generated by a linear function plus some noise:

$$\vec{y} = h_\beta(\mathrm{X}) + \vec{\epsilon}$$

Assumptions:
1. The relation between x (independent variable) and y (dependent variable) is linear in a parameter $\boldsymbol{\beta}$.
2. ϵ i are independent, identically distributed random variables following a N(0, σ2) distribution. (Note: σ is constant)

Least Squares:

$$\begin{aligned}\beta_{LS} &= \arg\min_{\beta} \sum_i d(h_\beta(x_i), y_i) \\ &= \arg\min_{\beta} \|\vec{y} - h_\beta(\mathrm{X})\|_2^2 \\ &= \arg\min_{\beta} \|\vec{y} - \beta\mathrm{X}\|_2^2\end{aligned}$$

An Unbiased Estimator:

$$\begin{aligned}E[\beta_{LS}] &= E[(X^TX)^{-1}X^Ty] \\ &= (X^TX)^{-1}X^TE[y] \\ &= (X^TX)^{-1}X^TE[X\beta + \epsilon] \\ &= (X^TX)^{-1}X^TX\beta + E[\epsilon] \\ &= \beta\end{aligned}$$

Logistic Regression:
Our goal is to fit a linear model to the log-odds of being in one of our classes

$$\log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = \alpha + \beta X$$

$$\begin{aligned}P(y_i = 1|x_i) &= \begin{cases} logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 1 \\ 1 - logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 0 \end{cases} \\ &= (logit^{-1}(\alpha + \beta x_i))^{y_i}(1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}\end{aligned}$$

So we can define:

$$L(\alpha, \beta) = \prod_i (logit^{-1}(\alpha + \beta x_i))^{y_i}(1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}$$

Z-values:
These are the number of standard deviations from the mean of a N(0,1) distribution required in order to contain a specific % of values were you to sample a large number of times.

Extending our Linear Model:
Changing the assumptions we made can drastically change the problem we are solving. A few ways to extend the linear model:
1. Non-constant variance - used in WLS (weighted least squares)
2. Distribution of error is not Normal - used in GLM (generalized linear models)