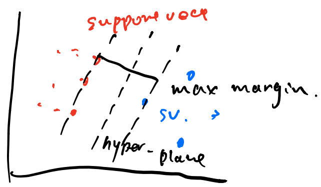
- They are both for classification.
- LR $\approx$ SVM with

Support Vector Machine.

- objective: Find the hyperplane that has the maximum margin in N-d feature space.



Logistic Regression:

- take the output of the linear func and map the value in [0, 1]. using sigmoid func (logistic func)
  probability.

Loss Function:

- SVM: $\min_w \lambda \|w\|^2 + \sum_i \max\{0, 1 - y_i w^T x_i\}$
- LR: $\min_w \lambda \|w\|^2 + \sum_i \log(1 + e^{-y_i w^T x_i})$

LR: $P(y=1|x) = \frac{1}{1+e^{-z}}$

$P(y=0|x) = \frac{e^{-z}}{1+e^{-z}}$

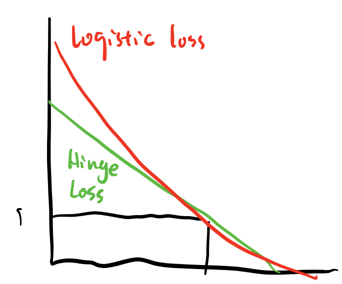
$J(w) = -\sum y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)})$: Cross-entropy

Likelihood:

$P(y|\pi) = \prod^n \pi_i^{y_i} (1-\pi_i)^{1-y_i}$

$\to L(\theta) = \prod^n P(x_i; \theta)$

$\log(L(\theta))$: Log-likelihood.



- Logistic loss diverge. sensitive to outliers.

Difference:

- SVM maximize the margin. LR not.
- LR more sensitive to outliers. The cost func of LR diverge fast
- LR gives probability SVM not.

General Advice:

1. Try LR first
2. If fails. & Data not linearly separable.
   → SVM with none-linear kernel

n # features m # samples.

- n large
  → LR / SVM (linear kernel)
- n small, m intermediate (n=1-1000, m=10-10,000)
  → SVM (Gaussian kernel)
- n small, m large (n=1-1000, m=50,000+)
  → add features. LR/ SVM (linear kernel).

- SVM find the widest possible separating margin.
  LR optimizes the log-likelihood, with probabilities modeled by sigmoid
- SVM extends by using kernel. transforming datasets into rich feature space
  complex data can be dealt with in the lifted hyper space.
- Both linearly separable.