Optimization Approach to Classification

	Training	Evaluation
Regression	minimize error (MSE)	MSE
Classification	minimize loss (cross entropy)	Accuracy/error.
↳ Logistic regression		
↳ Softmax regression		

We cannot minimize error (Optimization Approach) in Classification

instead, We minimize a surrogate loss (e.g. Cross Entropy)

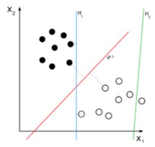
"替代损失"

1. Probabilistic Approach $P(y|\vec{x}) ? \rightarrow$ "$\hat{y} = \arg\max_y P(y|\vec{x})$"
2. Optimization Approach

"Where to place the decision boundary"

1. parametric method: $y = wx + b \Rightarrow$ Compute w, b

"SVM" Support Vector Machine.

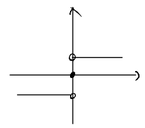


2. non-parametric method:

kNN (k-nearest neighbors), Decision Tree/ Random Forest

$D = \{x_i, y_i\}_{i=1}^N$ where $y_i \in \{+1, -1\}$.

$\hat{y} = \text{sign}(w^T x + b)$, where $\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$



$L(w,b) = \frac{1}{N}\sum_{i=1}^N L_{0/1}(y_i, \hat{y}_i)$

$L_{0/1}(y_i, \hat{y}_i) = \mathbb{1}(y_i \neq \hat{y}_i)$ "Zero/One Loss function"

Empirical Risk Minimization

↳ "Average loss on limited training set"

What we're using in regression:

Expected Risk Minimization

↳ "Expected Loss on true data distribution"

Problem: We can not minimize it using Gradient Descent!

Since Sign()'s gradient is 0 and is not continuous.

$\Rightarrow \text{sign}(w^T x + b) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$

$\Rightarrow$ Let $z = w^T x + b$.

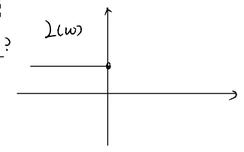
$L_{0/1}(y, \hat{y}) = \mathbb{1}(y \neq \hat{y}) = \mathbb{1}(\underbrace{y \cdot z}_{\text{"}\hat{y}\text{"}} \leq 0)$

$\therefore L(w,b) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}(y_i z_i \leq 0)$ $(z_i = w^T x_i + b)$

Gradient: for any weight w_j

$\frac{\partial L(w,b)}{\partial w_j} = \frac{1}{N}\sum_{i=1}^N \frac{\partial\, \mathbb{1}(y_i z_i \leq 0)}{\partial\, y_i z_i} \cdot \frac{\partial\, y_i z_i}{w_j}$, where $\frac{\partial\, \mathbb{1}(y_i z_i \leq 0)}{\partial\, y_i z_i} = 0$.

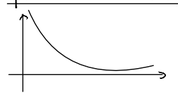
↳ Zero! Then How to proceed GD?



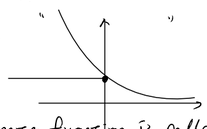
Convex Function

real-value function $f(x)$. satisfies:

$\forall t \in [0,1]. \forall x_1, x_2, x_1 < x_2$, $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$



We use a Convex function to approximate $L_{0/1}(w,b)$



The approximate function is called "Surrogate Loss"

$\Rightarrow$ minimize $L_{0/1} \Leftrightarrow$ minimize Surrogate loss.

Loss function for classification:

- **Zero/one loss**: $L_{0/1}(y,z) = \mathbb{1}(yz \leq 0)$.
- **Logistic loss**: $L_{log}(y,z) = \frac{1}{\log 2}\log(1+\exp(-yz))$ ≡ logistic regression
- **Hinge loss**: $L_{hin}(y,z) = \max\{0, 1-yz\}$ $L_{hin} + \lambda\|w\|_2^2$: SVM
- **Exponential loss**: $L_{exp}(y,z) = \exp(-yz)$
- **Squared loss**: $L_{sqr}(y,z) = (y-z)^2$

Loss function for regression:

- MSE
- Cross Entropy
- Huber Loss

Generative Approach for classification: Naive Bayes.

Optimization Approach:

$y = \text{sign}(w^T x + b) \Rightarrow$ Learn parameter by ↘ Surrogate loss

$L = \frac{1}{N}\sum_{i=1}^N L(y_i, z_i) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}(y_i z_i \leq 0)$

Probabilistic Approach:

$P(y|x,w) \Rightarrow$ Learn parameter by ↘ cross entropy

$L = -\frac{1}{N}\sum_{i=1}^N \log P(y_i|x_i, w)$

They tell us "How to predict y from x"

i.e. they don't care How x was generated by y

$\Rightarrow$ "Discriminative models" 判别式模型

Generative Approach:

$P(y)$: prior

$P(x|y)$: class conditional prob

$\Rightarrow P(y|x)$ what we want.

Assume $\begin{cases} x_i \sim P(x|y_i) \\ y_i \sim P(y) \end{cases} \Rightarrow D = \{x_i, y_i\}_{i=1}^N$

Generative Method: $-\frac{1}{N}\sum_{i=1}^N \log P(y_i, x_i|\theta)$	Minimizing Cross Entropy
Discriminative Method: $-\frac{1}{N}\sum_{i=1}^N \log P(y_i|x_i, \theta)$	

Naive Bayes:

$P(\vec{x}|y) = P(x_0|y) \cdot P(x_1|y) \cdots P(x_D|y)$ Assume independence

$P(y|x) = \frac{P(y)\,P(\vec{x}|y)}{P(x)}$

$\hat{y} = \arg\max_x P(y|\vec{x})$