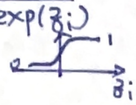
Linear Predictive Model

$$z_i = b^T x_i + b_0$$

$$P(y_i=1 \mid x_i) = \frac{\exp(z_i)}{1+\exp(z_i)}$$

sigmoid function



Softmax Function

$$\text{softmax}(z_i)_j = \frac{\exp(z_{ij})}{\sum_{l=1}^{J} \exp(z_{il})}$$

$$P(y_i=j \mid x_i) = \text{softmax}(z_i)$$

Empirical Risk Minimization

$$b^* = \arg\min_b \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \sigma(f(x_i;b)))$$

Loss Function

$$\ell(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y}) \quad [0,1]$$

$$\ell(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) \quad [0,1,2\ldots]$$

K-Nearest Neighbors

- classifier:
  1. set $K = 1, 2, \ldots$
  2. get x_i & $x_0, x_1, x_2 \ldots$ distance
  3. select K nearest neighbor
  4. majority voting
- regressor:
  1. set K
  2. get x_i & $x_0, x_1, x_2 \ldots$ distance
  3. $\hat{y} = \frac{1}{K} \sum_{i \in K_{nn}} y_i$

     or

     $$\hat{y} = \frac{\sum_i \frac{1}{d(x,x_i)} y_i}{\sum_i \frac{1}{d(x,x_i)}}$$

Overfitting ⟷ Underfitting

Train → Valid → Test

Cross Validation

- reduce variance of performance estimates
- N-fold cross validations

Metric

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Accuracy:

$$\frac{\#\text{ correct guesses}}{\#\text{ of data points}}$$

99% sensitive

└ 99% of (+) cases → predict (+)

99% specific

└ 99% of (−) cases → predict (−)

	(−) predict	(+) predict
(+) true	FN	TP
(−) true	TN	FP

$$\text{accuracy} = \frac{TN+TP}{\text{All}}$$

K-fold crossvalidation

```
for fold in folds:
    valid ← fold
    train ← folds \ fold
    for k in k_list
        train on train
        valid on valid
        get_performance
        store performance
evaluate score on each k
k* ← argmax_k score_avg(k)
```

feature selection ANOVA analysis of var

univariate

$$F\text{-score} = \frac{\text{explained var}}{\text{unexplained var}} = \frac{n_0(\bar{x}_0-\bar{X})^2 + n_1(\bar{x}_1-\bar{X})^2}{(n_0-1)\sigma_0^2 + (n_1-1)\sigma_1^2}$$

How different the group means are
> random noise

→ high: group means different; feature rich

→ low: " close; feature poor

sequential feature selection (forward)

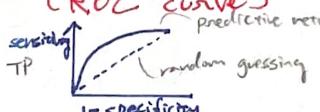
1. no feature
2. while model improve
   - add feature for f_i in f_list
   - evaluate
   - choose best feature f_i
   - add f_i^*

	univariate	v	sequential
PRO	fast easy interpretable		capture interaction accuracy ↗
CON	ignore correlation		greedy slow overfit

sequential feature selection (backward)

1. all features
2. while model improve
   - remove feature for f_i in f_list
   - remove f_i: doesn't
     is → improves/ hurt

Receiving Operating Characteristic (ROC curve)



AUC: integral of the plot under the curve

1. classifier output $p \in [0,1]$
2. vary threshold $0 \to 1$
   → let $p > \tau =$ (+)
3. for each τ, get TPR, FPR
4. plot TPR v FPR
   (sensitivity) (1−specificity)

Confusion Matrix

	Predict 0 1 2 … N
Actual 0 1 2 ⋮ N	

	+	−
+	TP	FN
−	FP	TN

$$TPR: \frac{TP}{TP+FN}$$

$$FPR: \frac{FP}{FP+TN}$$

p is the p^{th} feature

Standard Scaling (Z-scoring)

$$\tilde{x}_{ip} = \frac{x_{ip}-\hat{\mu}}{\hat{\sigma}_p}$$

$\tilde{\mu} = 0$

$\tilde{\sigma} = 1$

Min-Max Rescaling

$$\tilde{x}_{ip} = \frac{x_{ip} - \hat{x}_p^{min}}{\hat{x}_p^{max} - \hat{x}_p^{min}}$$

Max-abs Rescaling

$$\tilde{x}_{ip} = \frac{x_{ip}}{|\hat{x}|_p^{max}}$$

LASSO v RIDGE

LASSO	RIDGE
many w_i → 0	smoother
L1-norm	L2-norm
some f_i irrelevant use this	most f_i useful

∨ combine ‖ Elastic Net

SVM

1. $w^T x + b \geq m$ (•)

   $w^T x + b \leq -m$ (x)

   $$w^*, b^* = \arg\max_{w,b} m \quad \text{s.t. } \|w\|_2 = 1$$

2. $w^T x_i + b \geq m$

   $\to \frac{w^T}{m} x_i + \frac{b}{m} \geq 1$

3. $w^* = \arg\min_w \|w\|_2^2$

   $w^T x + b > 1 \quad \forall$ (•)

   $w^T x + b < -1 \quad \forall$ (x)

Slack variables

$$w^* = \arg\min_w \|w\|_2^2 + \lambda \sum_{i=1}^{n} \psi_i$$

Kernel

$x \to \phi(x)$

- linear
- polynomial
- exponential
- radial basis

decision tree

cross-entropy loss

$$L_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

$$\hat{p}_{mk} = \frac{\#\text{samples w/ outcome } k \text{ in node } m}{\#\text{ of samples in node } m}$$

random forest aiding decision tree

- smoother boundaries
- ensemble
- reduce variance
- increase bias

K-means

0. $SSE = \sum_{i=1}^{n} \|x_i - c_{z_i}\|_2^2$
1. given $x_1, x_2, \ldots x_n \in \mathbb{R}^d$ & cluster no. k
2. initialized $c_0, c_1, c_2 \ldots c_k$
3. repeat until convergence
   - assignment:
     for each data i, do:
     $z_i \leftarrow \arg\min_{j \in \{0 \ldots k\}} \|x_i - c_j\|_2^2$
     ↳ index of nearest centroid
   - centroid update
     $$c_j = \frac{1}{n_j}\sum_{i \in S_j} x_i$$

GMM

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

$$\pi_k \Rightarrow \sum_{k=1}^{K} \pi_k = 1$$

Expectation - Maximization Algo

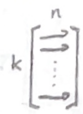
$$X = \{x_1, x_2, \ldots, x_N\}$$

↳ assume this is generated w/ a Gaussian Mixture Model

1. hidden variables r_{ik}
   ↳ how much does k^{th} distribution "own" datapoint x_i?
   $$r_{ik} = P(z_i = k \mid x_i)$$
2. initial guess
   $$\mu_k, \Sigma_k, \pi_k$$
3. Expectation step
   $$r_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$
   ↳ compute the relative prob. that x_i belongs to k^{th} distribution
4. Maximization step
   $$N_k = \sum_{i=1}^{N} r_{ik}$$
   $$\pi_k = N_k / N$$
   $$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} x_i$$
   $$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$
5. $\mathcal{L} = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$
6. loop → until $\mathcal{L}$ not improving

PCA

0. w/ n features, k data pts
$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \in \mathbb{R}^{k \times n}$$



1. $$\Sigma = \frac{1}{n-1} X^T X$$
   $$\Sigma \in \mathbb{R}^{n \times n}$$
2. $$\Sigma V_i = \lambda_i V_i \quad \in \mathbb{R}^n$$
   cov. / vec of / magnitude of that PC
   "principal components"
   $$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$
3. $W = [\, v_1 \; v_2 \ldots v_\ell \,] \in \mathbb{R}^{n \times \ell}$
   $\ell \leq n$
4. $\underset{\mathbb{R}^{k\times n}}{X} \cdot \underset{\mathbb{R}^{n\times\ell}}{W}$: $Z = X \cdot W \in \mathbb{R}^{k \times \ell}$
5. $\hat{X} = Z W^T$

t-SNE

t-distributed stochastic neighbor embedding

0. $X = \{x_1, x_2 \ldots x_n\}$
   perplexity — local size
   K — target dim
   lr, max-iter

1. for x_i in X:
   find σ_i, s.t. entropy of P_i = perplexity
   for $x_j \neq x_i$:
   $$p_{ij} = \exp(-\|x_i - x_j\|^2 / (2\sigma_i^2))$$
   $$p_{ij} = p_{ij} / \sum_{j'} p_{ij'}$$