Introduction to Machine Learning by Ethem Alpaydin

sample stat is a func computed from $X_1, \dots, X_n$ random var, thus itself is also a random var. $f(X_1, \dots, X_n)$

→ Sample mean → sample variance

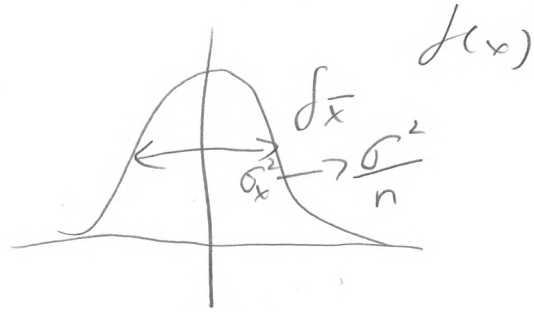
Assume $X_1, \dots X_n$ is iid (independent, identically distributed)

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \qquad Var(X) = E[(x - \bar{x})^2] = E[X^2] - (\bar{x})^2$$

mean of sample distribution of means $= \mu_{\bar{x}}$

Thm.1 $E(\bar{x}) = \mu_{\bar{x}} = \mu$

mean of stat — mean of pop.



Thm.2 $E[(X-\mu)^2] = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

variance of sampling mean distribution = variance of population

infinite pop. / finite pop. w/ replacement

Thm.3 $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ finite pop. w/o replacement

N sample size n

Thm.4 If pop. is ~~uniform~~ normally dist. w/ μ, σ^2, sample should dist. w/ μ & $\frac{\sigma^2}{n}$

Thm.5 If pop. has μ, σ^2 but not necessarily normal, standardized variable associated w/ $\bar{x}$, given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$ $\bar{X}$: sampling mean

is asymptotically normal, $\lim_{n\to\infty} P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-w^2/2}\, dw$

Thm 3. ☐ from central limit then.
Otherwise thm 6 will still be correct if replace $\frac{\sigma}{\sqrt{n}}$ by $\sigma_{\bar{x}}^2$ as given in thm 3

Sampling Binomial Prop

$\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$\therefore \mu = p, \ \sigma = \sqrt{\frac{pq}{n}}$

Since $0 \le p \le 1$,

$E(x) = 0 \cdot q + 1 \cdot p = p$

$E(x^2) = 0^2 \cdot q + 1^2 \cdot p = p$

$p - p^2 = p(1-p) = pq$

Sampling variance

$s^2 = \frac{(x_1 - \bar{x})^2 + \dots (x_n - \bar{x})^2}{n}$

unbiased estimator

$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{\sum_{k}^{n} (x_k - \bar{x})^2}{n-1}$

On Ave. right on target

Smaller Var gives more efficient estimator.

Bessel's correction: replace n by $(n-1)$

$\mu \pm n\sigma$ (68-95-99 rule) (~~confidence interval~~)

$\mu \pm 1.96\sigma$: 95% $\mu \pm 2.58\sigma$ (0.99) crit. values

(confidence ~~interval~~ levels) z_c

$n \ge 30$ $\begin{cases} \bar{X} \pm z_c \sigma_{\bar{x}} \text{ or } \pm 2.58 \sigma_{\bar{x}} & (\text{w/ replace / } \infty \text{ pop.}) \\ \bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} & (\text{w/o replacement}) \end{cases}$

$n < 30$: t-distribution $\pm t_{0.975}$ s: s.d.

95%: $-t_{0.975} < \frac{(\bar{x} - \mu)\sqrt{n}}{s} < t_{0.975}$

Master Thm.

$T(1) = O(1)$

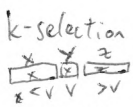
$T(n) = O(n^r) + T(\lceil \frac{n}{\beta} \rceil)$

$\log_\beta \alpha < r \Rightarrow O(n^r)$

$\log_\beta \alpha = r \Rightarrow O(n^r \log_\beta n)$

$\log_\beta \alpha > r \Rightarrow O(n^{\log_\beta \alpha})$

k-selection



$x < v$ | v | $x > v$

$x = k$-th return v

$<$ recurse lft

$>$ recurse rht

choose v in $\leq O(n)$

$x \geq \frac{n}{2}$ $z \geq \frac{n}{2}$

Dominance Counting

1. sort $\ln$ such that each side $\lceil \frac{n}{2} \rceil$ pt.s by k-selection $O(n)$
2. find pts dominated on lft, rht recursively
3. sort lft and rht pts using y-axis
4. merge sort inversion

Strassen

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} P_5+P_4-P_2+P_6 & P_1+P_2 \\ P_3+P_4 & P_1+P_3-P_2-P_7 \end{bmatrix}$$

$P_1 = A_{11}(B_{12}-B_{22})$ $P_2 = (A_{11}+A_{12})B_{22}$

$P_3 = (A_{21}+A_{22})B_{11}$ $P_4 = A_{22}(B_{21}-B_{11})$

$P_5 = (A_{11}+A_{22})(B_{11}+B_{22})$ $P_6 = (A_{12}-A_{22})(B_{21}+B_{22})$

$P_7 = (A_{11}-A_{21})(B_{11}+B_{12})$

pf $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ $\hat{p} \approx p$

Error type I: reject hypothesis H_0 (false neg)

- Level of confidence ≈ 0.05 or 0.01 use S.D.

complement of S.D.

Type II: False neg False pos.

95% confident of z-score between -1.96 and 1.96

If z-score outside mostly 0.05 prob.

1-tailed / 2-tailed

Level of significance

0.10	0.05	0.01	0.005	
± 1.28	± 1.645	± 2.33	± 2.58	1-tailed
± 1.645	± 1.96	± 2.58	± 2.81	2-tailed

null: H_0 gives $\mu = u$ assume $u = 12$

H_1 $\mu > u$ $\mu < u$ $n = 36$

p values $P(z \geq 1.9) = 0.029$ $P(z \leq 1.9)$ $\bar{x} = 12.98$

$\therefore$ about $\frac{3}{100}$ $z = 1.9$

that $\bar{x} \geq 12.98$

$\mu \neq u = P(z \geq 1.9) + P(z \leq 1.9)$

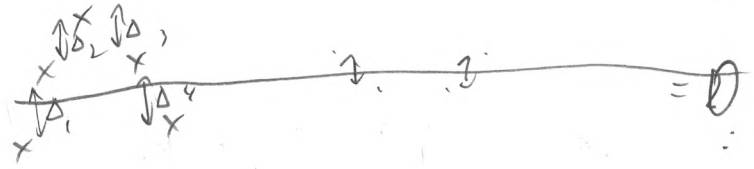
Small p values suggests rejecting H_0

Method of least squares $\underbrace{d_1^2 + \dots + d_n^2}_{\text{minimum}}$ small

Linear approx.

$$\sum_{j=1}^{n} y_i = an + b \sum_{j=1}^{n} x_j$$

$$\sum_{j=1}^{n} x_j y_i = a \sum_{j=1}^{n} + b \sum_{j=1}^{n} x_j^2$$



$$a = \frac{\sum y (\sum x^2) - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \qquad b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

3 curves

$$Y, Y^{-1}, 1 - g(y)$$

$$\underbrace{X}_{\text{var}} = \{x^t, r^t\}_{t=1}^{N}$$

data pt. no.

Data Collecting Amount:

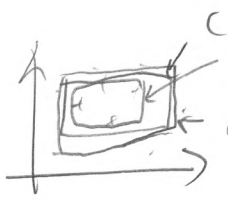
Probably approximately correct (PAC)

VC Dimension (relate learner complexity to errors)

class C: class to classify $\rightarrow E(h|X) = \sum_{t=1}^{N} 1(h(x^t) \neq r^t)$

Error (Total Err)

$h(x) = \begin{cases} 1 & \text{if classifys as pos} \\ 0 & \text{if classifys as neg} \end{cases}$



S: most specific hypothesis

G: most general hypothesis

C: classify region

$h \in H$, between ~~H~~ S & G ∃ consistent & make up the version space

PAC — How many N training examples such that w/ prob $(1-\delta)$ h has error at most ε?

$P\{C \Delta h \leq \varepsilon\} \geq 1-\delta$ (Δh, region between C and h

VC Dim — Capacity of learning machine

N pts can be labeled in 2^N for shattered

If $h \in H$ separates, H shatters N

~~Any~~ Max no. of N pts h shatters.

look for test err w/ train err

↑ cap → ↓ performance over-fitting

$$\text{test error} \leq \text{training err} + \sqrt{\frac{C\left(\log\left(\frac{2N}{C}+1\right) - \log\left(\frac{\eta}{4}\right)\right)}{N}}$$

N: num. of samples

C: VC dim

penalty (complexity)

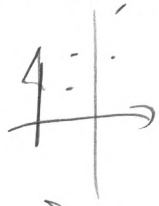
VC Dim for linear

In n-dim feature space set of $n+1$ ($n > n$) if in ~~shattered~~ general dimension IFF no subset of $(n+1)$ pts lies on $(n-1)$ hyperplane

↑ ⟹ no $n+1$ pts. lies on $n-1$ hyperplane

$\mathcal{H}$ shatters n pts in n-dim if all possible combinations of n pts in n-dim space correctly classified by $\mathcal{H}$

$\mathcal{H}$, not h, shatters, classifier, not actual hypothesis



s.t. lin. shatters 3 pts in 2 dim

VC dim = cardinality of largest set of pts $\mathcal{H}$ can shatter

VC dim linear: $(n+1)$ n: dim

Shatterly: $\exists h \in \mathcal{H} : E(h, X) = 0$

Noise, Latent Var

$$\hat{r}_{ij} = \sum_{k=1}^{K} p_{ik} q_{kj} \quad \text{for } i = 1 \dots |U| \ \& \ j = 1 \dots |D|$$

$$e_{ij}^2 = (r_{ij} - \hat{r}_{ij})^2 = \left(r_{ij} - \sum_{k=1}^{K} p_{ik} q_{kj}\right)^2$$

$$\therefore \frac{\partial e_{ij}^2}{\partial p_{ik}} \qquad \frac{\partial e_{ij}^2}{\partial q_{kj}}$$

$$= -2 e_{ij} q_{kj} \qquad = -2 e_{ij} p_{ik}$$

Gradient descent $\eta \approx 0.002$

Minimize for observed ratings only $\rightarrow \hat{R}$

$$\sum_{T} e_{ij}^2 = \sum_{r_{ij} \in T} \left(r_{ij} - \sum_{k=1}^{K} p_{ik} q_{kj}\right)^2$$

$$e_{ij}^2 = \left(r_{ij} - \sum_{k=1}^{K} p_{ik} q_{kj}\right)^2 + \frac{\beta}{2}\left(\sum_{a=1}^{|U|} \sum_{b=1}^{K} (p_{ab})^2 + \sum_{a=1}^{K} \sum_{b=1}^{|D|} (q_{ab})^2\right)$$

Matrix Factorization

- make recommendation under partial information

Let A be $m \times n$ and of rank r

$A = X Y \quad X : m \times r \quad Y \; r \times n$

$\therefore$ storage saving for by

$A = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 6 & 10 \\ 3 & 6 & 7 & 13 \end{bmatrix} \quad r = 2$

$A_1 = 1 \cdot A_1 + 0 A_3 \quad A_3 = 0 A_1 + 1 A_3$

$A_2 = 2 \cdot A_1 + 0 A_3 \quad A_4 = 2 A_1 + 1 A_3$

$\therefore X = \begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 3 & 7 \end{bmatrix} \quad Y = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$

e.g. netflix

$U = |U|$ users $\quad D = |D|$ movies $\therefore = A \; (|U| \times |D|)$

Rating $: R \; (|U| \text{ by } |D|)$

$\therefore \quad R = P Q$ ## relates each user w/ some latent features

$R \approx P \times Q = \hat{R}$

M-step: $\Phi^{l+1} = \arg\max_{\Phi} Q(\Phi|\Phi^l) = \sum_t \sum_i h_i^t [\log \pi_i + \log p_i(x^t|\Phi^l)]$

$= \arg\max \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p_i(x^t|\Phi^l)$

$\pi_i = \frac{\sum h_i^t}{N}$ $\quad \nabla_{\pi_i} \sum_t \sum_i h_i^t \log \pi_i - \lambda \sum_i \pi_i - 1 = 0$

$\nabla_\Phi \sum_t \sum_i h_i^t \log p_i(x^t|\Phi) = 0$

If Gaussian $\hat{p}(x^t|\Phi) \sim N(m_i, S_i)$ M-Step $\Rightarrow$

$m_i^{l+1} = \frac{\sum_t h_i^t x^t}{\sum_t h_i^t}$ $\quad S_i^{l+1} = \frac{\sum_t h_i^t (x^t - m_i^{l+1})(x^t - m_i^{l+1})^T}{\sum_t h_i^t}$

$h_i^t = \frac{\pi_i |S_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x^t - m_i)^T S_i^{-1}(x^t - m_i)\right)}{\sum_j \pi_j |S_j|^{-\frac{1}{2}} \exp\left(\frac{1}{2}(x^t - m_j)^T S_j^{-1}(x^t - m_j)\right)}$

Hierarchical Clustering → Agglomerative (N → k Group) / Divisive (1 → k)

Minkowski $d_m(x^r, x^s) = \left(\sum_j^d (x_j^r - x_j^s)^p\right)^{1/p}$ euclidean for $p = 2$

City-block $d_{cb}(x^r, x^s) = \sum_j^d |x_j^r - x_j^s|$ $p=1 \rightarrow$ manhattan

single-link = min $d(\,)$ complete link = max $d(\,)$ Ave-/link = ave $d(\,)$ centroid $x^r \in G_i, x^s \in G_j$

Non-param est.

Histogram $= \frac{|x^t \text{ in bin}|}{Nh}$ Naive $= \frac{|x^t \text{ in range}|}{Nh}$

Kernel func $K(u) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{u^2}{2}\right)}$ kernel est $= \frac{1}{Nh} \sum_{t=1}^N K\left(\frac{x - x^t}{h}\right)$

↳ kNN: $\hat{p}(x) = \frac{k}{2N d_k(x)}$

generalization on multivariate Gaussian: $K(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{\left(-\frac{\|u\|^2}{2}\right)}$

$\hat{p}(x) = \frac{1}{Nh^d} \sum_{t=1}^N K\left(\frac{x - x^t}{h}\right)$

curse of dimensionality

Classification

$\hat{p}(x|C_i) = \frac{1}{Nh^d} \sum_{t=1}^N K\left(\frac{x - x^t}{h}\right) r_i^t \rightarrow g_i(x) = \hat{p}(x|C_i)\hat{P}(C_i)$

knn $\hat{p}(x|C_i) = \frac{k_i \leftarrow \text{no. of neighbours} \in C_i \text{ in } k\text{-ell}}{N_i V^k(x) \leftarrow \text{volume of } x^k \text{ from } x}$ $\rightarrow P(C_i|x) = \frac{k_i}{k}$

Condensed nearest neighbour → 1-nn

For all $x \in X$

Find $x' \in Z$ | $|x - x'| = \min_{x^i \in Z} |x - x^i|$

If class(x) ≠ class(x') add to Z

Until Z does not change (end)

Distance based

$$D(x, m_i) = \min_{j=1}^{k} (x, m_j) \qquad D(x, m_i) = |x - m_i| \sqrt{(x-m_i)^T S_i^{-1} (x-m_i)}$$

or Mahalanobis $D(x, x^t | M) = (x - x^t)^T M (x - x^t)$

$M = L^T L$ $\quad$ L: k×d, M: d×d

$$= |z - z^t|^2 (Lx - Lx^t)^T (Lx - Lx^t)$$
$$= |z - z^t|^2$$

Outlier detection

$$LOF(x) = \frac{d_k(x)}{\sum_{s \in N(x)} d_k(s) / |N(x)|}$$ ← number of samples

> 1 → outlier

Non-parametric regression

Same as classification, where r^t no longer classifier, but result

regression → $\hat{g}(x) = \frac{\sum b(x, x^t) r^t}{\sum b(x, x^t)}$ $\quad b(x, x^t)$ if x^t is in x's

running-mean

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right)}$$

$$w(u) = \begin{cases} 1 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

kernel

$$\hat{g}(x) = \frac{\sum_t K\left(\frac{x - x^t}{h}\right) r^t}{\sum_t K\left(\frac{x - x^t}{h}\right)}$$

k, h small → ↑ complexity ↑ var ↓ bias

k, h big → ↓ complexity ↓ var ↑ bias

decision tree — leave: result — inland decision nodes: (binary) Numeric-split / Discrete split

greedy algorithm

recursive split

Univariate

$\hat{P}(C_i|x,m) = p_m^i = \frac{N_m^i}{N_m}$ N_m = Number of instances at node m

node m pure if $p_m^i = 0$ or 1, impurity $I_m = -\sum_i^K p_m^i \log_2 p_m^i$ (entropy)

$\Phi = -p \log_2 p - (1-p)\log_2(1-p)$ (entropy) $[0,1]$ $=0 \to$ pure

$\Phi(p,1-p) = 2p(1-p)$ (Gini index) $\Phi(p,1-p) = 1 - \max(p,1-p)$ (misclassification err.)

If node is pure, generate leaf and stop, else split recursively

Impurity after split: N_{mj} of N_m branch j

$\hat{P}(C_i|x,m,j) = p'^i_{mj} = \frac{N'^i_{mj}}{N_{mj}}$ $I'_m = -\sum_{j=1}^{n} \frac{N_{mj}}{N_m} \sum_{i=1}^{K} p'^i_{mj} \log_2 p'^i_{mj}$

To generate tree:

```
if entropy < θ1
    create leaf of majority
    return
else
    split
    for all branches,
        generate tree(branch)
```

Split: for all attributes $i = 1 \dots d$

if x_i is discrete of n vals. (discrete)

e = split entropy $(X_1 \dots X_n)$

if $e <$ min, best $\leftarrow i$, min $\leftarrow e$

else (Numeric)

split X into X_1, X_2 on x_i

$e \leftarrow$ split entropy

if $e <$ Min, min $\leftarrow e$, best $\leftarrow i$

return best

Univariate regression $X_{mj} \subset X_m$ of taking branch j

$$b_{mj}(x) = \begin{cases} 1 & \text{if } x \in X_{mj} \\ 0 & \end{cases}$$

$$g_{mj} = \frac{\sum_t b_{mj}(x^t) r^t}{\sum_t b_{mj}(x^t)} \qquad E'_m = \frac{1}{N_m} \sum_j \sum_t (r^t - g_{mj})^2 b_{mj}(x^t)$$

minimize error E'_m, replace entropy with E'_m, θ_l & θ_r

Pre-pruning: early stop Post-pruning: grow tree then prune what overfitted on pruning set

Rule learning

tree induction (BFS) rule induction (DFS)

1. rule set contains rules
2. remove covered (true) samples of rules in ruleset when added

Sequential, adds 1 rule at a time till all evaluated.

Multivariate,

$$\text{Node}_m = f_m(x) : \underbrace{w_m^T x + w_{m0} > 0}_{\text{linear}} \quad \text{hyperplane}$$

For discrete attributes should be 0/1 dummy numeric

Divide until defined by polyhedra in input space

- or $f_m(x) = x^T W_m x \; w_m^T x + w_{m0} > 0$ → quadratic
- or $f_m(x) = |x - c_m| \le \alpha_m$ sphere node

Discriminant Assume model for $g_i(x|\Phi_i)$ and find boundary

Linear: $g_i(x|w_i, w_{i0}) = w_i^T x + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0}$ $O(d)$

when shared cov., and almost linearly separable

Quad: $g_i(x|W_i, w_i, w_{i0}) = x^T W_i x + w_i^T x + w_{i0}$ $O(d^2)$

or convert to linear

$g_i(x) = \sum_{j=1}^{k} w_j \phi_{ij}(x)$ → basis func: $\sin(x_1)$, $\exp(-(x_1 - m)^2/c)$ $(g(x)\ I(x>c)\ldots$

$K=2$: $g(x) = g_1(x) - g_2(x) = w^T x - w_0$

$r = \frac{g(x)}{|w|}$ $r_0 = \frac{w_0}{|w|}$ find w using LDA

$K>2$ $g_i(x|w_i, w_{i0}) = w_i^T x + w_{i0} = \begin{cases} \geq 0 & \text{if } x \in C_i \\ \leq 0 & \text{otherwise} \end{cases}$

$\max_k g_j(x)$ $|g_i(x)|/|w_i|$ is dist

Pairwise Separation Use $K(K-1)/2$ linear discriminants, $g_{ij}(x)$

where $g_{ij}(x|w_{ij}, w_{ij0}) = w_{ij}^T x + w_{ij0}$

$p(C_1|x) = \text{sigmoid}(w^T x + w_0) = \frac{1}{1+\exp(-(w^T x + w_0))}$