Linear Algebra

vector & scalar properties:
$x+y = y+x$
$a0 = 0$ (zero vector 0)
$0x = 0$
$a(x+y) = ax + ay$
Distance $= \|x-y\|$ (norm)

Norms

Euclidean "2-Norm"
$g(x) = \sqrt{\sum_{d=1}^{D} x_d^2}$ "birds eye diag"
$\|x\|$ or $\|x\|_2$
$x^T x = \|x\|^2$
$x^T y \leq \|x\|\|y\|$ Cauchy Schwarz inequality

Manhattan Norm "grid" "ℓ_1-norm"
$g(x) = \sum_{d=1}^{D} |x_d|$
Max Norm $- \|x\|_\infty$
$g(x) = \max_d |x_d|$
Zero Norm
$g(x) = \sum_{d=1}^{D} \mathbb{1}[x_d \neq 0]$
adds up non-zeros

Vector dot product

$x \cdot y = \sum_{d=1}^{D} x_d y_d$
commutativity
$x \cdot y = y \cdot x$
distributivity
$x \cdot (y+z) = x \cdot y + x \cdot z$

Orthogonal = perpendicular when $x \cdot y = 0$

Matrices

- element-wise product $\begin{smallmatrix}1&2&3\\4&5&6\end{smallmatrix} * \begin{smallmatrix}1&2&3\\4&5&6\end{smallmatrix} = \begin{smallmatrix}1&4&9\\16&25&36\end{smallmatrix}$
- inner product / dot product

$(AB)_{n,m} = \sum_{k=1}^{K} A_{n,k} B_{k,m}$ (11 12 / 21 22 / 31 32)

Transpose $\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}^T = \begin{bmatrix}1&2\\3&4\\5&6\end{bmatrix}$
$(A^T)_{m,n} = A_{n,m}$
$A^T A = A$
$A = A^T$ if square & symmetric

associative $A(BC) = (AB)C$
distributive $A(B+C) = AB + AC$
NOT commutative $AB \neq BA$
$I = \begin{bmatrix}1&0\\0&1\end{bmatrix}$ $I^T = I$ & $IA = AI = A$

Determinant
$\begin{bmatrix}a&b\\c&d\end{bmatrix} = ad - bc = |A|$

Inverse
$A^{-1} = \frac{1}{|A|}\begin{bmatrix}d&-b\\-c&a\end{bmatrix} = \frac{1}{ad-bc}\begin{bmatrix}d&-b\\-c&a\end{bmatrix}$

commutativity: $AA^{-1} = A^{-1}A = I$
scalar multi: $(aA)^{-1} = a^{-1}A^{-1}$ $a \neq 0$
transposition: $(A^T)^{-1} = (A^{-1})^T$
product: $(AB)^{-1} = B^{-1}A^{-1}$
diagonal $\mathrm{diag}\begin{pmatrix}1&1&0\\1&2&0\\0&1&3\end{pmatrix} = \langle 1,2,3 \rangle = \begin{bmatrix}1&0&0\\0&2&0\\0&0&3\end{bmatrix}$ diag matrix

Differential Calculus

derivative is the slope of function at that point
$\frac{\partial f}{\partial x}$ often noted f'
to find extreme point, take derivative & set to 0. $\frac{\partial f}{\partial x}\ 2x^2 - 3x - 1 = \partial_x f(x) = 4x - 3$
$4x - 3 = 0 \quad x = 3/4$

Rules
scalar multip. $\partial_x [a f(x)] = a[\partial_x f(x)]$ polynomial: $\partial_x [x^k] = kx^{k-1}$
function addition: $\partial_x [f(x) + g(x)] = [\partial_x f(x)] + [\partial_x g(x)]$
function multip: $\partial_x [f(x) g(x)] = f(x)[\partial_x g(x)] + [\partial_x f(x)] g(x)$
function division: $\partial_x \left[\frac{f(x)}{g(x)}\right] = \frac{[\partial_x f(x)] g(x) - f(x)[\partial_x g(x)]}{[g(x)]^2}$
exponentiation: $\partial_x [e^x] = e^x$ and $\partial_x [a^x] = \log(a) e^x$
logarithms: $\partial_x [\log x] = \frac{1}{x}$

Integral Calculus

= computing area under curve
$\int_a^b dx\, f(x)$ = area under curve between a & b
can "kind of" read integral as $\sum_{x=a}^{b} f(x)$

Logarithms

quotient rule: $\log\left(\frac{x}{y}\right) = \log(x) \cdot \log(y)$
product rule: $\log(x)(y) = \log(x) + \log(y)$
reciprocal rule: $\log\left(\frac{1}{x}\right) = -\log(x)$
power rule: $\log(x^y) = y \cdot \log(x)$
properties: log neg # = undefined
log of 0 = undefined $\quad \log(1) = 0$
$\log(\infty) = \infty \quad \log(e) = 1 \quad \log(e^x) = x \quad e^{\log(x)} = x$
*log is natural log (i.e. ln)

precision: of predicted true, a high fraction is true, but some true may be predicted false
recall: high recall do not miss true examples, if predicted false highly likely to be false but some predicted true may be false

Probability

$P(\text{heads}) = \theta$
$P(\text{tails}) = 1 - \theta$
Bernoulli: $P(x) = \theta^x (1-\theta)^{1-x}$, $x \in \{0,1\}$
complement: $P(\bar{x}) = 1 - P(x)$
conditional: $P(x|y) = \frac{P(x,y)}{P(y)}$ $\quad P(x|y,z) = \frac{P(x,y,z)}{P(y,z)}$
$P(x,y|z) = \frac{P(x,y,z)}{P(z)}$

product rule: (AND) aka joint probability *also $P(X \cap Y)$
$P(x,y) = P(x) \cdot P(y|x)$ *dependent
$P(x,y,z) = P(x) \cdot P(y|x) \cdot P(z|x,y)$
$= P(x) \cdot \left[\frac{P(x,y)}{P(x)}\right] \cdot \left[\frac{P(x,y,z)}{P(x,y)}\right]$

independent: $P(x,y) = P(x) \cdot P(y)$
mutually exclusive: $P(x,y) = 0$

addition / union (or) *mutually exclusive
$P(X \cup Y) = P(X) + P(Y) - P(X,Y)$ $\quad P(X \cup Y) = P(X) + P(Y)$

Independence $(X \perp Y)$ if:
$P(X|Y) = P(X)$
$P(Y|X) = P(Y)$
$P(X,Y) = P(X)P(Y)$

Conditional Independence: $(X \perp Y | Z)$
$P(x|y,z) = P(x|z)$
$P(y|x,z) = P(y|z)$
$P(x,y|z) = P(x|z)P(y|z)$

BAYES THEOREM

$$P(C|x) = \frac{P(C)P(x|C)}{P(x)} = \frac{P(C_i)P(x|C_i)}{\sum_{k=1}^{K} P(x|C_k)P(C_k)}$$

(prior: $P(C)$; likelihood: $P(x|C)$; posterior: $P(C|x)$; evidence: $P(x)$)

for $K=2$ $P(x) = P(x|C=1)P(C=1) + P(x|C=0)P(C=0)$

for something like $P(X_1=1, X_2=0, X_3=1 | Y=0)$
$P(X_1=1, X_2=0, X_3=1, Y=0)$
$P(x) = P(X_1=1, X_2=0, X_3=1 | Y=1) + P(X_1=1, X_2=0, X_3=1 | Y=0)$
↑ both for $Y=0$ & $Y=1$

posterior: the probability that outcome C_i occurs given some conditions X
prior: what is known regarding possible values that C_i might take before looking at samp or how many times C_i occurred independently of conditions (x)
Likelihood: conditional probability that an event belonging to class C_i has observation X
Evidence: # of times outcome X occurred

for bag-o-words
$P(Y|X_{1:d}) = \frac{P(X_{1:d}|Y)P(Y)}{P(X_{1:d})} = \frac{''}{\sum_{y'} P(X_{1:d}|Y')P(y')}$
joint probability

Losses & Risk

α_i = action i
Loss of α_i when state C_k: λ_{ik}

	C_1	C_2	$\cdots$	C_K
α_1	0	λ_{12}		λ_{1K}
α_2	λ_{21}	0		
$\vdots$			$\ddots$	$\vdots$
α_i	λ_{i1}		$\cdots$	0

$\lambda_{ik} = \begin{cases} 0 & \text{if } i=k \\ 1 & \text{otherwise} \end{cases}$

$R(\alpha_i|x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|x) = \sum_{k \neq i} P(C_k|x) = 1 - P(C_i|x)$ ← weighted average

choose α_i if $R(\alpha_i|x) = \min_k R(\alpha_k|x) = \max_i P(C_i|x)$

$\min_i \frac{1}{K} \sum_{k=1}^{K} \lambda_{ik}$ ← average

Reject

$\lambda_{ik} = \begin{cases} 0 & \text{if } i=k \\ \lambda & \text{if } i=K+1,\ 0<\lambda<1 \\ 1 & \text{otherwise} \end{cases}$

$R(\alpha_{K+1}|x) = \sum_{k=1}^{K} \lambda P(C_k|x) = \lambda$
$R(\alpha_i|x) = \sum_{k \neq i} P(C_k|x) = 1 - P(C_i|x)$

	C_1	C_2	$\cdots$	C_K
α_1	0	1		1
α_2	1	0		1
$\vdots$				0
α_K	1			
α_{K+1}	λ	$\lambda \cdots$		λ

Discriminant function

maps input variables to class label
Choose C_i if $g_i(x) = \max_k g_k(x)$
$g_i(x): \begin{cases} -R(\alpha_i|x) \\ P(C_i|x) \\ p(x|C_i)P(C_i) \end{cases}$
K decision regions $R_1 \ldots R_K$
$R_i = \{x \mid g_i(x) = \max_k g_k(x)\}$
Dichotomizer ($K=2$)
$g(x) = g_1(x) - g_2(x)$
choose $\begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$
Log odds $\log \frac{P(C_1|x)}{P(C_2|x)}$
polychotomizer $K>2$

parametric classification
$g_i(x) = p(x|C_i)P(C_i) \propto P(C_i|x)$
$g_i(x) = \log p(x|C_i) + \log P(C_i)$
Gaussian
$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu)^2}{2\sigma_i^2}\right]$
$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$

ML estimates
$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$ $\quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$
$s_i^2 = \frac{\sum_t (x^t - m)^2 r_i^t}{\sum_t r_i^t}$
boundary $g_1 \ldots$

Parametric

Estimator – algorithm ($d(x)$) to predict i.e. $\frac{1}{N}\sum_{i=1}^{N} x_i$

Estimators: treat θ as random variable w/ prior $p(\theta)$

Bayes rule $p(\theta|X) = p(X|\theta)p(\theta)/p(X)$

MAP – Maximum a priori:

$\theta_{MAP} = \text{argmax}_\theta\, p(\theta|X) = \text{argmax}_\theta\, p(X|\theta)p(\theta)$ * } point estimate

Maximum Likelihood (ML):

$\theta_{ML} = \text{argmax}_\theta\, p(X|\theta)$ * } point estimate

w/ infinite data MLE + MAP will converge → find MODE

Bayes: $\theta_{MAP} = \theta_{Bayes}$ for Gaussian

$\theta_{Bayes} = E[\theta|X] = \int \theta p(\theta|X) d\theta$ * makes predictions for new data

* finds MEAN

parametric: $X = \{x^t\}$ where $x^t \sim (p(x))$

assume a form for $p(x|\theta)$ + estimate θ + stats using X e.g. $N(\mu, \sigma^2)$ where $\theta = \{\mu, \sigma^2\}$

Likelihood of θ given sample X (θ assumed random var)

$l(\theta|X) = p(X|\theta) = \prod_t p(x^t|\theta)$ * conditional independence

Log Likelihood * bc of underflow problem

$\mathcal{L}(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)$

Maximum likelihood estimator (MLE)

$\theta^* = \text{argmax}_\theta\, \mathcal{L}(\theta|X)$

① Calculate likelihood:

for Bernoulli: $f_x(x) = \lambda^x(1-\lambda)^{1-x}$

with sample $X = \{x_1, x_2, x_3\}$ $x_1 = 1$, $x_2 = 1$, $x_3 = 0$

$L(\lambda) = P(X|\lambda) = P(x_1|\lambda) \cdot P(x_2|\lambda) \cdot P(x_3|\lambda)$

$= \lambda^1(1-\lambda)^{1-1} \cdot \lambda^1(1-\lambda)^{1-1} \cdot \lambda^0(1-\lambda)^{1-0}$

$= \lambda^2(1-\lambda)$

② Calculate log likelihood

$\log L(\lambda) = \log(\lambda^2(1-\lambda))$

$= 2\log(\lambda) - \log(\lambda - 1)$

* can take derivative of likelihood or log likelihood?

③ For MLE:

take derivative + set to 0

$\mathcal{L}'(\lambda) = 2\lambda - 3\lambda^2 = 0$

$\lambda = 0, \; 2/3$ (circled) because 0 not in interval

② Find MAP

- take likelihood + given prior dist

$p(\theta|X) = p(\theta) \cdot L(\theta)$ (prior, likelihood)

- plug in prior values
- take log
- take derivative + set to 0

Bayes "frequentist estimator"

Full dist $= P(X_5 | HTHHH) = P(X_5 | 75\%)$

$= \begin{cases} 75\% & X_5 = H \\ 25\% & X_5 = T \end{cases}$

Full: $p(x|X) = \int p(x|\theta)p(\theta|X)d\theta$

example: $x^t \sim N(\theta, \sigma_0^2)$

$+ \theta \sim N(\mu, \sigma^2)$

$\theta_{ML} = m = \sum x^t / N$

$\theta_{MAP} = \theta_{BAYES} =$

$E[\theta|X] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2}\mu$

sample mean → (first term) → 1 for large samples; (second term) → 0 prior

$E_X[d_X] = E\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] = \mu$

*basically after integration $E_X[d_X]$ = mean

Bias $b_\theta(d) = E_X[d(X)] - \theta$

Variance: $E_X[(d(X) - E_X[d(X)])^2]$

Mean square error: $r(d, \theta) = E[(d-\theta)^2]$

$= \text{bias}^2 + \text{variance}$

↓ bias ↓ variance: Estimator d is $\frac{1}{N}\sum x_i$ or $\sqrt{\frac{1}{a}\sum x_i^2}$

↓ bias ↑ variance: just $\sum_{i=1}^{N} x_i$ (not normalized)

↑ bias ↓ variance $\frac{1}{N}\sum x_i + E \cdot 10^9$ or constantly output 0

↑ bias ↑ variance $\frac{1}{N}\sum_{i=1}^{N} x_i + 10^9 + E \cdot 10^9$, $E \sim N(0,1)$ or sample from uniformly at random huge range x^i

Distributions:

Bernoulli: 2 states x in $\{0,1\}$

$p(x|p_0) = p_0^x(1-p_0)^{(1-x)}$

$\mathcal{L}(p_0|X) = \log \prod_t p_0^{x^t}(1-p_0)^{(1-x^t)}$

MLE: $p_0 = \frac{1}{N}\sum_t x^t$

Multinomial $K>2$ states x_i in $\{0,1\}$

$P(x_1, x_2 \cdots x_K | p_1, p_2 \cdots p_K) = \prod_i p_i^{x_i}$

$\mathcal{L}(p_1, p_2, p_K | X) = \log \prod_t \prod_i p_i^{x_i^t}$

MLE: $p_i = \frac{1}{N}\sum_t x_i^t$ * must be non-neg + add to one

Gaussian (Normal) Dist

$p(x|\theta) = N[\mu, \sigma^2]$ $\theta = (\mu, \sigma^2)$

$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

MLE for μ + σ^2

$m = \frac{\sum_t x^t}{N}$ $s^2 = \frac{\sum_t (x^t - m)^2}{N}$

Dirichlet – multinomial w/ $K>2$ states

sample likelihood $p(X|q) = \prod_{t=1}^{N}\prod_{i=1}^{K} q_i$ states $q = [q_1, q_2 \cdots q_K]^T$

$\sum_i q_i = 1$

prior q

$\text{Dirichlet}(q|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{i=1}^{K} q_i^{\alpha_i - 1}$

where $\alpha = [\alpha_1, \cdots, \alpha_K]^T$ + $\alpha_0 = \sum_i \alpha_i$

$\Gamma(x)$ = gamma $= \int_0^\infty u^{x-1}e^{-u}du$

* conjugate prior

posterior $p(q|X) \propto \prod q_i^{\alpha_i + N_i - 1}$ = Dirichlet $(q|\alpha + n)$

where $n = [N_1, \cdots N_K]^T$ + $\sum_i N_i = N$

mean $E[q_i] = \frac{\alpha_i}{\sum_k \alpha_k}$

Mode: $q_i = \frac{\alpha_i - 1}{\sum_{j=1}^{K}\alpha_j - K}$, $\alpha_i > 1$

Beta variable is binary $x^t \in \{0,1\}$ becomes Bernoulli

$p(X|q) = \prod_t q^{x^t}(1-q)^{1-x^t}$

+ dirichlet prior reduces to beta:

$\text{beta}(q|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$

* conjugate prior

posterior: $p(q|A, N, \alpha, \beta) \propto q^{A+\alpha-1}(1-p)^{N-A+\beta-1}$

* when $\alpha = \beta = 1$ uniform + posterior same shape as likelihood

mean $= \frac{\alpha}{\alpha+\beta}$

mode $= \frac{\alpha - 1}{\alpha+\beta-2}$

Naive Bayes assume features ind. given class

$P(X_{1:d}|y) = \prod_{j=1:d} P(x_j|y)$

Params = $(|X| - 1)$