

Notation: $P_{X|Y}(x \mid y) = P[X = x \mid Y = y]$
*Subscript indicates the RV, and the value indicates the realization.
Intro:
Random Experiment: An outcome for each run.
Sample Space Ω : Set of all possible outcomes.
Event: Measurable subsets of Ω .
Prob. of Event A: $P(A) = \frac{\text{Number of outcomes in A}}{\text{Number of outcomes in } \Omega}$
Axioms: (1) $P(A) \geq 0 \forall A \in \Omega$, (2) $P(\Omega) = 1$,
(3) If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B) \forall A, B \in \Omega$
Cond. Prob. $P(A|B) = \frac{P(A \cap B)}{P(B)}$
*Prob. measured on new sample space B .
 $*P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
Independence: $P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$
Total Prob. Thm: If H_1, H_2, \dots, H_n form a partition of Ω , then $P(A) = \sum_{i=1}^n P(A|H_i)P(H_i)$.
Bayes' Rule: $P(H_k|A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(A|H_k)P(H_k)}{\sum_{i=1}^n P(A|H_i)P(H_i)}$
*Posteriori: $P(H_k|A)$, Likelihood: $P(A|H_k)$, Prior: $P(H_k)$
1 RV:
Cumulative Distribution Fn (CDF): $F_X(x) = P[X \leq x]$
Prob. Mass Fn (PMF): $P_X(x_j) = P[X = x_j] \ j = 1, 2, \dots$
Prob. Density Fn (PDF): $f_X(x) = \frac{d}{dx} F_X(x)$
 $*P[a \leq X \leq b] = \int_a^b f_X(x) \, dx$
Exp.: $E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) \, dx$
 $E[h(X)] = \sum_{k=-\infty}^{\infty} h(k) P_X(x_i = k)$
Variance: $\sigma_X^2 = \text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
Cond. Exp.: $E[X|A] = \int_{-\infty}^{\infty} x f_X(x|A) \, dx$
2 RVs:
Joint PMF: $P_{X,Y}(x, y) = P[X = x, Y = y]$
Joint PDF: $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$
 $*P[(X, Y) \in A] = \int \int_{(x,y) \in A} f_{X,Y}(x, y) \, dx \, dy$
Exp.: $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) \, dx \, dy$
Correlation: $E[XY]$
Covar.: $\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$
Corr. Coeff.: $\rho_{X,Y} = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y}$
 $*-1 \leq \rho_{X,Y} \leq 1$
Marginal PMF: $P_X(x) = \sum_{j=1}^{\infty} P_{X,Y}(x, y_j) \mid P_Y(y)$
Marginal PDF: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \mid f_Y(y)$
Cond. PMF: $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} \mid P_{Y|X}(y|x)$
Cond. PDF: $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \mid f_{Y|X}(y|x)$
Bayes' Rule
 $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y') \, dy'}$
 $*P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{j=1}^{\infty} P_{X|Y}(x|y_j)P_Y(y_j)}$
Ind.: $f_{X|Y}(x|y) = f_X(x) \forall y \Leftrightarrow f_{X,Y}(x, y) = f_X(x)f_Y(y)$
Thm: If independent, then uncorrelated unless Gaussian.
Uncorrelated: $\text{Cov}[X, Y] = 0 \Leftrightarrow \rho_{X,Y} = 0$
Orthogonal: $E[XY] = 0$
Cond. Exp.: $E[Y] = E[E[Y|X]]$ or $E[E[h(Y)|X]]$
 $*E[E[Y|X]]$ w.r.t. $X \mid E[Y|X]$ w.r.t. Y .
Estimation: Estimate unknown parameter θ from n i.i.d. measurements X_1, X_2, \dots, X_n , $\hat{\Theta}(\underline{X}) = g(X_1, X_2, \dots, X_n)$
Estimation Error: $\hat{\Theta}(\underline{X}) - \theta$.
Unbiased: $\hat{\Theta}(\underline{X})$ is unbiased if $E[\hat{\Theta}(\underline{X})] = \theta$.
Asymptotically unbiased: $\lim_{n \rightarrow \infty} E[\hat{\Theta}(\underline{X})] = \theta$.
Consistent: $\hat{\Theta}(\underline{X})$ is consistent if $\hat{\Theta}(\underline{X}) \rightarrow \theta$ as $n \rightarrow \infty$ or $\forall \epsilon > 0, \lim_{n \rightarrow \infty} P[|\hat{\Theta}(\underline{X}) - \theta| < \epsilon] \rightarrow 1$.
Sufficient: A statistic is sufficient if the expression depends only on the statistic, it should be made up of x_1, x_2, \dots, x_n .
Sample Mean: $M_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^n X_i$.
*Given a sequence of i.i.d. RVs, X_1, X_2, \dots, X_n , M_n is unbiased and consistent.
Sample Variance: $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M_n)^2$.
*Given a sequence of i.i.d. RVs, X_1, X_2, \dots, X_n , S_n^2 is biased and consistent.
*Use $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_n)^2$ for unbiased.
Chebychev's Inequality: $P[|X - E[X]| \geq \epsilon] \leq \frac{\text{Var}[X]}{\epsilon^2}$
 $*P[|X - E[X]| < \epsilon] \geq 1 - \frac{\text{Var}[X]}{\epsilon^2}$
Weak Law of Large #s: $\lim_{n \rightarrow \infty} P[|M_n - \mu| < \epsilon] = 1 \forall \epsilon > 0$.
ML Estimation: Choose θ that is most likely to generate the obs. x_1, x_2, \dots, x_n .
*Disc: $\hat{\Theta} = \arg \max_{\theta} P_{\underline{X}}(\underline{x}|\theta) \stackrel{\log}{=} \hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log P_X(x_i|\theta)$
*Cont: $\hat{\Theta} = \arg \max_{\theta} f_{\underline{X}}(\underline{x}|\theta) \stackrel{\log}{=} \hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i|\theta)$
Maximum A Posteriori (MAP) Estimation:
*Disc: $\hat{\theta} = \arg \max_{\theta} P_{\Theta|\underline{X}}(\theta|\underline{x}) = \arg \max_{\theta} P_{\underline{X}|\Theta}(\underline{x}|\theta)P_{\Theta}(\theta)$
*Cont: $\hat{\theta} = \arg \max_{\theta} f_{\Theta|\underline{X}}(\theta|\underline{x}) = \arg \max_{\theta} f_{\underline{X}|\Theta}(\underline{x}|\theta)f_{\Theta}(\theta)$
 $*f_{\Theta|\underline{X}}(\theta|\underline{x})$: Posteriori, $f_{\underline{X}|\Theta}(\underline{x}|\theta)$: Likelihood, $f_{\Theta}(\theta)$: Prior
Bayes' Rule: $P_{\Theta|\underline{X}}(\theta|\underline{x}) = \begin{cases} \frac{P_{\underline{X}|\Theta}(\underline{x}|\theta)P_{\Theta}(\theta)}{P_{\underline{X}}(\underline{x})} & \text{if } \underline{X} \text{ disc.} \\ \frac{f_{\underline{X}|\Theta}(\underline{x}|\theta)f_{\Theta}(\theta)}{f_{\underline{X}}(\underline{x})} & \text{if } \underline{X} \text{ cont.} \end{cases}$
 $f_{\Theta|\underline{X}}(\theta|\underline{x}) = \begin{cases} \frac{P_{\underline{X}|\Theta}(\underline{x}|\theta)f_{\Theta}(\theta)}{P_{\underline{X}}(\underline{x})} & \text{if } \underline{X} \text{ disc.} \\ \frac{f_{\underline{X}|\Theta}(\underline{x}|\theta)f_{\Theta}(\theta)}{f_{\underline{X}}(\underline{x})} & \text{if } \underline{X} \text{ cont.} \end{cases}$
*Independent of θ : $f_{\underline{X}}(\underline{x}) = \int_{-\infty}^{\infty} f_{\underline{X}|\Theta}(\underline{x}|\theta)f_{\Theta}(\theta) \, d\theta$
Least Mean Squares (LMS) Estimation: Assume prior $P_{\Theta}(\theta)$ or $f_{\Theta}(\theta)$ w/ obs. $\underline{X} = \underline{x}$.
 $*\hat{\theta} = g(\underline{x}) = E[\Theta|\underline{X} = \underline{x}] \mid \hat{\Theta} = g(\underline{X}) = E[\Theta|\underline{X}]$
Conditional Exp. $E[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) \, dx$
Binary Hyp. Testing: H_0 : Null Hyp., H_1 : Alt. Hyp.
 $\Omega_{\underline{X}}$: Set of all possible obs. \underline{x} .



TI Err. (False Rejection): Reject H_0 when H_0 is true.

* $\alpha(R) = P[\underline{X} \in R \mid H_0]$ (false alarm)

TII Err. (False Accept.): Accept H_0 when H_1 is true.

* $\beta(R) = P[\underline{X} \in R^c \mid H_1]$ (missed detection)

Likelihood Ratio Test: $\forall \underline{x} \quad L(\underline{x}) = \frac{P_{\underline{X}}(\underline{x} \mid H_1)}{P_{\underline{X}}(\underline{x} \mid H_0)} \underset{H_0}{\gtrless} 1$ or ξ

***Max. Likelihood Test:** 1, **Likelihood Ratio Test:** ξ

Neyman-Pearson Lemma: Given a false rejection prob. (α), the LRT offers the smallest possible false accept. prob. (β), and vice versa.

*LRT produces (α, β) pairs that lie on the efficient frontier.



Bayesian Hyp. Testing:

MAP Rule: $L(\underline{x}) = \frac{p_{\underline{X}}(\underline{x} \mid H_1)}{p_{\underline{X}}(\underline{x} \mid H_0)} \underset{H_0}{\gtrless} \frac{P[H_0]}{P[H_1]}$

Gaussian to Q Fcn: 1. Find $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$.

2. Use table to find $Q(x)$ for $x \geq 0$.

Min. Cost Bayes' Dec. Rule: C_{ij} is cost of choosing H_j when H_i is true. Given obs. $\underline{X} = \underline{x}$, the expected cost of choosing H_j is $A_j(\underline{x}) = \sum_{i=0}^1 C_{ij} P[H_i \mid \underline{X} = \underline{x}]$.

Min. Cost Dec. Rule: $L(\underline{x}) = \frac{P_{\underline{X}}(\underline{x} \mid H_1)}{P_{\underline{X}}(\underline{x} \mid H_0)} \underset{H_0}{\gtrless} \frac{(C_{01} - C_{00})P[H_0]}{(C_{10} - C_{11})P[H_1]}$.

* C_{01} : False accept. cost, C_{10} : False reject. cost.

Naive Bayes Assumption: Assume X_1, \dots, X_n (features) are ind., then $p_{\underline{X}}(\underline{x} \mid \theta) = \prod_{i=1}^n p_{X_i}(x_i \mid \theta)$.

Notation: $P_{\underline{X}}(\underline{x} \mid \theta)$, only put RVs in subscript, not values.

$P_{\underline{X}}(\underline{x} \mid H_i)$, didn't put H in subscript b/c it's not a RV.

Binomial # of successes in n trials, each w/ prob. p

$b(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, \dots$

* $E[X] = \mu = np$ | $\text{Var}(X) = \sigma^2 = np(1-p)$

Multinomial # of x_i successes in n trials, each w/ prob. p_i

$f(x_i \mid p_i \forall i, n) = \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}$

* $\sum_i x_i = n$, and $\sum_{i=1}^m p_i = 1$

* $E[X_i] = \mu_i = np_i$ | $\text{Var}(X_i) = \sigma_i^2 = np_i(1-p_i)$

Hypergeometric # of successes in n draws from N items, k of which are successes

$h(x \mid N, n, k) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$

* $\max\{0, n - (N - k)\} \leq x \leq \min\{n, k\}$

* $E[X] = \mu = \frac{n k}{N}$ | $\text{Var}(X) = \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N} \cdot \left(1 - \frac{k}{N}\right)$

Negative Binomial # of trials until k successes, each w/ prob. p

$b^*(x \mid k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$

* $x \geq k$, $x = k, k+1, \dots$

* $E[X] = \mu = \frac{k}{p}$ | $\text{Var}(X) = \sigma^2 = \frac{k(1-p)}{p^2}$

Geometric # of trials until 1st success, each w/ prob. p

$g(x \mid p) = p(1-p)^{x-1}$

* $x \geq 1$, $x = 1, 2, 3, \dots$

* $E[X] = \mu = \frac{1}{p}$ | $\text{Var}(X) = \sigma^2 = \frac{1-p}{p^2}$

Poisson # of events in a fixed interval w/ rate λ

$p(x \mid \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$

* $x \geq 0$, $x = 0, 1, 2, \dots$

* $E[X] = \mu = \lambda t$ | $\text{Var}(X) = \sigma^2 = \lambda t$

Beta Prior Θ is a Beta R.V. w/ $\alpha, \beta > 0$

$f_{\Theta}(\theta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$

* $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$

Prop.: 1. $\Gamma(x+1) = x\Gamma(x)$. For $m \in \mathbb{Z}^+$, $\Gamma(m+1) = m!$.

2. $\beta(\alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} = \beta \binom{\alpha+\beta-1}{\alpha-1}$

3. Expected Value: $E[\Theta] = \frac{\alpha}{\alpha+\beta}$ for $\alpha, \beta > 0$

4. Mode (max of PDF): $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$

Drawing Beta Dist. 1. Label x -axis from 0 to 1. 2. Identify mode.

3. Determine shape based on α and β : $\alpha = \beta = 1$ (uniform), $\alpha = \beta > 1$ (bell-shaped, peak at 0.5), $\alpha = \beta < 1$ (U-shaped w/ high density near 0 and 1), $\alpha > \beta$ (left-skewed), $\alpha < \beta$ (right-skewed).

Uniform PDF $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

* $E[X] = \frac{a+b}{2}$, $\text{Var}[X] = \frac{(b-a)^2}{12}$

Random Vector: $\underline{X} = (X_1, \dots, X_n) = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = [X_1 \quad \dots \quad X_n]^T$

Mean Vector: $\underline{m}_X = E[\underline{X}] = [\mu_1, \dots, \mu_n]^T$

Corr. Mat.: $R_X = \begin{bmatrix} E[X_1^2] & \dots & E[X_1 X_n] \\ E[X_2 X_1] & \dots & E[X_2 X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] & \dots & E[X_n^2] \end{bmatrix}$

*Real, symmetric ($R = R^T$), and PSD ($\forall \underline{a}, \underline{a}^T R \underline{a} \geq 0$).

Covar. Mat.: $K_X = \begin{bmatrix} \text{Var}[X_1] & \dots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_2, X_1] & \dots & \text{Cov}[X_2, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \dots & \text{Var}[X_n] \end{bmatrix}$

* $K_X = R_X - \underline{m}_X \underline{m}_X^T$

*Diagonal $K_X \iff X_1, \dots, X_n$ are (mutually) uncorrelated.

Lin. Trans. $\underline{Y} = A\underline{X}$ (A rotates and stretches \underline{X})

Mean: $E[\underline{Y}] = A\underline{m}_X$

Covar. Mat.: $K_Y = AK_XA^T$

Diagonalization of Covar. Mat. (Uncorrelated):

$\forall \underline{X}$, set $P = [\underline{e}_1, \dots, \underline{e}_n]$ of K_X , if $\underline{Y} = P^T \underline{X}$, then

$K_Y = P^T K_X P = \Lambda$

* \underline{Y} : Uncorrelated RVs, $K_X = PAP^T$

Find an Uncorrelated K_Y

1. Find eigenvalues, normalized eigenvectors of K_X .

2. Set $K_Y = \Lambda$, where $\underline{Y} = P^T \underline{X}$

PDF of L.T. If $\underline{Y} = A\underline{X}$ w/ A not singular, then

$f_Y(\underline{y}) = \frac{f_X(\underline{x})}{|\det A|} \Big|_{\underline{x}=A^{-1}\underline{y}}$

Find $f_Y(\underline{y})$ 1. Given $f_X(\underline{x})$ and RV relations, define A

2. Determine $|\det A|$, A^{-1} , then $f_Y(\underline{y})$.

Gaussian RVs: $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$

PDF of jointly Gaus. $X_1, \dots, X_n \equiv$ Guas. vector:

$f_X(\underline{x}) = \frac{1}{(2\pi)^{n/2} |\det \Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1}(\underline{x}-\underline{\mu})}$

*1D: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

* $\underline{\mu} = \underline{m}_X$, $\Sigma = K_X$ (Σ not singular)

*Indep.: $f_X(\underline{x}) = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}$

*IID: $f_X(\underline{x}) = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$

*Cond. PDF: $f_X|\underline{Y}(\underline{x}|\underline{y}) = \mathcal{N}(\underline{\mu}_X|\underline{Y}, \Sigma_X|\underline{Y})$

Properties of Gaussian Vector:

1. PDF is completely determined by $\underline{\mu}$, Σ .

2. \underline{X} uncorrelated $\iff \underline{X}$ independent.

3. Any L.T. $\underline{Y} = A\underline{X}$ is Gaus. vector w/ $\underline{\mu}_Y = A\underline{\mu}_X$, $\Sigma_Y = A\Sigma_X A^T$.

4. Any subset of $\{X_i\}$ are jointly Gaus.

5. Any cond. PDF of a subset of $\{X_i\}$ given the other elements is Gaus.

Diagonalization of Guassian Covar. (Indep.)

$\forall \underline{X}$, set $P = [\underline{e}_1, \dots, \underline{e}_n]$ of Σ_X , if $\underline{Y} = P^T \underline{X}$, then

$\Sigma_Y = P^T \Sigma_X P = \Lambda$

* \underline{Y} : Indep. Gaussian RVs, $\Sigma_X = PAP^T$

How to go from Y to X ? 1. Given, $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$

2. $\underline{Y} \sim \mathcal{N}(\underline{0}, I)$ 3. $\underline{W} = \sqrt{\Lambda} \underline{V}$ 4. $\underline{Y} = P\underline{W}$ 4. $\underline{X} = \underline{Y} + \underline{\mu}$

Guassian Discriminant Analysis:

Obs: $\underline{X} = \underline{x} = (x_1, \dots, x_D)$

Hyp: \underline{x} is generated by $\mathcal{N}(\underline{\mu}_c, \Sigma_c)$, $c \in C$

Dec: Which "Gaussian bump" generated \underline{x} ?

Prior: $P[C = c] = \pi_c$ (Gaussian Mixture Model)

MAP: $\hat{c} = \arg \max_c P_C[c|\underline{X} = \underline{x}] = \arg \max_c f_{\underline{X}|C}(\underline{x} | c) \pi_c$

LGD: Given $\Sigma_c = \Sigma \forall c$, find c w/ best $\underline{\mu}_c$

$\hat{c} = \arg \max_c \underline{\beta}_c^T \underline{x} + \gamma_c$

* $\underline{\beta}_c^T = \underline{\mu}_c^T \Sigma^{-1} \mid \gamma_c = \log \pi_c - \frac{1}{2} \underline{\mu}_c^T \Sigma^{-1} \underline{\mu}_c$

Bin. Hyp. Decision Boundary $\underline{\beta}_0^T \underline{x} + \gamma_0 = \underline{\beta}_1^T \underline{x} + \gamma_1$

*Linear in space of \underline{x}

QGD: Given Σ_c are diff., find c w/ best $\underline{\mu}_c$, Σ_c

$\hat{c} = \arg \max_c -\frac{1}{2} \log |\Sigma_c| - \frac{1}{2} (\underline{x} - \underline{\mu}_c)^T \Sigma_c^{-1} (\underline{x} - \underline{\mu}_c) + \log \pi_c$

Bin. Hyp. Decision Boundary Quadratic in space of \underline{x}

How to find $\underline{\beta}_c$, $\underline{\mu}_c$, Σ_c : Given n points gen. by GMM, then n_c points $\{\underline{x}_1^c, \dots, \underline{x}_{n_c}^c\}$ come from $\mathcal{N}(\underline{\mu}_c, \Sigma_c)$

$\hat{\pi}_c = \frac{n_c}{n}$ (categorical RV)

$\hat{\mu}_c = \frac{1}{n_c} \sum_{i=1}^{n_c} \underline{x}_i^c$, (sample mean)

$\Sigma_c = \frac{1}{n_c} \sum_{i=1}^{n_c} (x_i^c - \hat{\mu}_c)(x_i^c - \hat{\mu}_c)^T$ (biased sampled var.)

Gaussian Estimation:

MAP Estimator for \underline{X} Given \underline{Y} When $\underline{W}=(\underline{X}, \underline{Y}) \sim \mathcal{N}(\underline{\mu}, \Sigma)$

Given $\underline{X} = \{X_1, \dots, X_n\}$, $\underline{Y} = \{Y_1, \dots, Y_m\}$

$\hat{\underline{\mu}}_{\text{MAP}}(\underline{y}) = \hat{\underline{\mu}}_{\text{LMS}}(\underline{y}) = \underline{\mu}_X|\underline{Y} = \underline{\mu}_X + \Sigma_{XY} \Sigma_{YY}^{-1} (\underline{y} - \underline{\mu}_Y)$

* $\hat{\underline{\mu}}_{\text{MAP/LMS}}$: Linear fcn of \underline{y}

Covar. Matrices: $\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$

* $\Sigma_{XX} = \Sigma_X = E \left[(\underline{X} - \underline{\mu}_X)(\underline{X} - \underline{\mu}_X)^T \right] \mid \Sigma_{YY} = \Sigma_Y$

* $\Sigma_{XY} = E \left[(\underline{X} - \underline{\mu}_X)(\underline{Y} - \underline{\mu}_Y)^T \right] \mid \Sigma_{YX} = \Sigma_{XY}^T$

Prec. Matrices: $\Lambda = \Sigma^{-1}$

Mean and Covar. Mat. of \underline{X} Given \underline{Y} :

* $\underline{\mu}_X|\underline{Y} = \underline{\mu}_X + \Sigma_{XY} \Sigma_{YY}^{-1} (\underline{y} - \underline{\mu}_Y)$

* $\Sigma_{X|\underline{Y}} = \Sigma_X - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$

***Reducing Uncertainty:** 2nd term is PSD, so given $\underline{Y} = \underline{y}$, always reducing uncertainty in \underline{X} .

ML Estimator for θ w/ Indep. Guas:

Given $\underline{X}=\{X_1, \dots, X_n\}$: $\hat{\theta}_{\text{ML}} = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}$ (weighted avg. \underline{x})

* $X_i=\theta + Z_i$: Measurement $\mid Z_i \sim \mathcal{N}(0, \sigma_i^2)$: Noise (indep.)

* $\frac{1}{\sigma_i^2}$: Precision of X_i (i.e. weight)

*Larger $\sigma_i^2 \implies$ less weight on X_i (less reliable measurement)

***SC:** If $\sigma_i^2 = \sigma^2 \forall i$ (iid), then $\hat{\theta}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$.

MAP Estimator for θ w/ Indep. Gaus., Gaus. Prior:

Given $\underline{X}=\{X_1, \dots, X_n\}$, prior $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$

$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} = \frac{1}{\sigma_0^2 + \sum_{i=1}^n \frac{1}{\sigma_i^2}} x_0 + \frac{1}{\sigma_0^2 + \sum_{i=1}^n \frac{1}{\sigma_i^2}} \hat{\theta}_{\text{ML}}$

* $X_i=\theta + Z_i$: Measurement $\mid Z_i \sim \mathcal{N}(0, \sigma_i^2)$: Noise (indep.)

* f_Θ : Gaussian prior \equiv prior meas. x_0 w/ σ_0^2 .

***SC:** As $n \rightarrow \infty$, $\hat{\theta}_{\text{MAP}} \rightarrow \hat{\theta}_{\text{ML}}$. As $\sigma_0^2 \rightarrow \infty$, $\hat{\theta}_{\text{MAP}} \rightarrow \hat{\theta}_{\text{ML}}$

LMMSE Estimator for \underline{X} Given \underline{Y} w/ non-Guas. \underline{X} , \underline{Y} :

$\hat{\underline{\mu}}_{\text{LMMSE}}(\underline{y}) = \underline{\mu}_X + \Sigma_{XY} \Sigma_{YY}^{-1} (\underline{y} - \underline{\mu}_Y)$

Linear Gaussian System: Given $\underline{Y} = A\underline{X} + \underline{b} + \underline{Z}$

* $\underline{X} \sim \mathcal{N}(\underline{\mu}_X, \Sigma_X)$, $\underline{Z} \sim \mathcal{N}(\underline{0}, \Sigma_Z)$: Noise (indep. of \underline{x})

* $A\underline{X} + \underline{b}$: channel distortion, \underline{Y} : Observed sig.

MAP/LMS Estimator for \underline{X} Given \underline{Y} w/ $\underline{W} = (\underline{X}, \underline{Y})$

Given $\underline{W} = \begin{bmatrix} \underline{X} \\ A\underline{X} + \underline{b} + \underline{Z} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \underline{X} \\ \underline{Z} \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{b} \end{bmatrix}$

$\hat{\mu}_{\text{MAP/LMS}} = \underline{\mu}_{\underline{X}} + \Sigma_{\underline{X}} A^T (A \Sigma_{\underline{X}} A^T + \Sigma_{\underline{Z}})^{-1} (\underline{y} - A \underline{\mu}_{\underline{X}} - \underline{b})$

$\Sigma_{\underline{X}\underline{Y}} = \Sigma_{\underline{X}} A^T, \Sigma_{\underline{Y}\underline{Y}} = A \Sigma_{\underline{X}} A^T + \Sigma_{\underline{Z}}$

$\hat{\mu}_{\text{MAP/LMS}} = \left(\Sigma_{\underline{X}}^{-1} + A^T \Sigma_{\underline{Z}}^{-1} A \right)^{-1} \left(A^T \Sigma_{\underline{Z}}^{-1} (\underline{y} - \underline{b}) + \Sigma_{\underline{X}}^{-1} \underline{\mu}_{\underline{X}} \right)$

***Use:** Good to use when \underline{Z} is indep.

Covar. Mat of \underline{X} Given $\underline{Y} = \underline{y}$: $\Sigma_{\underline{X}|\underline{y}} = \left(\Sigma_{\underline{X}}^{-1} + A^T \Sigma_{\underline{Z}}^{-1} A \right)^{-1}$

Linear Regression: Estimate unknown target fn $Y = g(\underline{X})$ w/ iid obs. $\{(\underline{x}_1, y_1), \dots, (\underline{x}_n, y_n)\}$ (MLE/MAP)

* $\underline{y} = [y_1 \dots y_n]^T$

* $\underline{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times D}$

ML Estimator: $Y = h(\underline{x}) = \underline{w}^T \underline{x} + Z \approx g(\underline{X})$, then $\hat{\underline{w}}_{\text{ML}} = (X X^T)^{-1} X^T \underline{y}$

*Assume $X^T X$ has full rank (i.e. invertible) since $n \gg D$

* n : # of obs., D : # of features.

* $\underline{x} = \{x_1, \dots, x_D\}$: Input features

* $\underline{w} = \{w_1, \dots, w_D\}$: Weights (parameter)

* $Z \sim \mathcal{N}(0, \sigma^2)$: Noise (i.i.d.)

* \underline{Y} : Target/observed output

* $X^\dagger = (X^T X)^{-1} X^T$: Pseudo-inverse of X (minimizes $\|X \underline{w} - \underline{y}\|_2^2 \iff$ maximizes the likelihood of training data)

Non-Linear Trans: $\hat{y} = \underline{w}^T \phi(\underline{x}) + Z$ w/ same assumptions,

then $\hat{\underline{w}}_{\text{ML}} = (X X^T)^{-1} X^T \underline{y}$

* $\phi(\underline{x})$: Non-linear transformation of \underline{x}

-E.g. of 1 dim x : $\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^M \end{bmatrix}$: Polynomial regression

* M : Degree of polynomial, $D = 1 + M$: # of features.

* $\underline{X} = \begin{bmatrix} \phi(\underline{x}_1)^T \\ \vdots \\ \phi(\underline{x}_n)^T \end{bmatrix} \in \mathbb{R}^{n \times D}$

Underfitting vs. Overfitting:

*Underfitting: Model too simple, high bias, low variance.

-Results in high train/test error.

*Overfitting: Model too complex, low bias, high variance.

-Results in low train error, high test error.

MAP Estimator (Bayesian Linear Regression): Assume

prior $w_i \sim \mathcal{N}(0, \tau^2)$ (i.i.d.) and $\hat{y} = \underline{w}^T \underline{x} + Z$, then

$\hat{\underline{w}}_{\text{MAP}} = (X^T X + \lambda I)^{-1} X^T \underline{y}$

* $\lambda = \frac{\sigma^2}{\tau^2}$: Regularization parameter

* \underline{X} : Can be linear or non-linear transformation of \underline{x}

* $\underline{x} = \{x_1, \dots, x_D\}$: Input features

* $\underline{w} = \{w_1, \dots, w_D\}$: Weights (parameter)

* $Z \sim \mathcal{N}(0, \sigma^2)$: Noise (i.i.d.)

* \underline{Y} : Target/observed output

Notes:

1. Useful when training data set size is small i.e. $n \ll D$.

2. Regularization: Prevents overfitting by penalizing large weights.

* $\tau = \infty \implies \lambda = 0$: No regularization so $\hat{\underline{w}}_{\text{MAP}} = \hat{\underline{w}}_{\text{ML}}$

* $\tau = 0 \implies \lambda = \infty$: Infinite regularization so $\hat{\underline{w}}_{\text{MAP}} = \underline{0}$

* $\tau \downarrow \implies \lambda \uparrow$: More regularization, simpler model.

* $\tau \uparrow \implies \lambda \downarrow$: Less regularization, more complex model.

Gaussian Linear System Given training data $\underline{Y} = \underline{X}\underline{w} + \underline{Z}$

$\hat{\underline{w}}_{\text{MAP}} = \underline{\mu}_{\underline{w}|\underline{Y}} = (X^T X + \lambda I)^{-1} X^T \underline{y}$

* $\underline{w} \sim \mathcal{N}(\underline{0}, \tau^2 I)$, $\underline{Z} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$

* $E[\hat{\underline{w}}(\underline{Y})] \rightarrow \underline{w}$ as $n \rightarrow \infty$

*Note: Matching it to canonical form.

Covar. Mat: $\Sigma_{\underline{w}|\underline{y}} = \left(\frac{1}{\sigma^2} X^T X + \frac{1}{\tau^2} I \right)^{-1} \preceq \tau^2 I$

-Less uncertainty in \underline{w} w/ more data. As $n \uparrow$, $\Sigma_{\underline{w}|\underline{y}} \downarrow$

Bayesian Prediction Given some new \underline{x}' (test data sample), find its label y'

Plug-In Approx: $\hat{Y}' = \underline{x}'^T \hat{\underline{w}}_{\text{MAP}}(\mathcal{D}) + Z'$

* \mathcal{D} : Training data set, $Z' \sim \mathcal{N}(0, \sigma^2)$: Noise

Bayesian Prediction: Use $Y' = \underline{x}'^T \underline{w} + Z'$ and

$f_{\underline{w}|\underline{Y}}(\underline{w} | \underline{y}) = \mathcal{N}(\underline{\mu}_{\underline{w}|\underline{Y}}, \Sigma_{\underline{w}|\underline{Y}})$ to return $f_{Y'}(y' | \mathcal{D})$ where

Y' is Gaussian given \mathcal{D} w/

* $\mu_{Y'|\mathcal{D}} = \underline{x}'^T \underline{\mu}_{\underline{w}|\underline{Y}}$

* $\sigma_{Y'|\mathcal{D}}^2 = \underline{x}'^T \Sigma_{\underline{w}|\underline{Y}} \underline{x}' + \sigma^2$

Linear Classification (Hyp. Test):

Binary Logistic Regression: Estimate \underline{w} s.t. it is a soft decision

$P_{Y|\underline{X}}(1 | \underline{x}) = \frac{P_{\underline{X}|\underline{Y}}(\underline{x}|1) P_Y(1)}{P_{\underline{X}|\underline{Y}}(\underline{x}|0) P_Y(0) + P_{\underline{X}|\underline{Y}}(\underline{x}|1) P_Y(1)}$

$P_{Y|\underline{X}}(1 | \underline{x}) = \frac{1}{1 + e^{-\alpha}} = \sigma(\alpha)$

* $P_{Y|\underline{X}}(0 | \underline{X}) = 1 - \sigma(\alpha) = \frac{1}{1 + e^{\alpha}} = \sigma(-\alpha)$

* $\alpha = \log \frac{P_{\underline{X}|\underline{Y}}(\underline{x}|1) P_Y(1)}{P_{\underline{X}|\underline{Y}}(\underline{x}|0) P_Y(0)} = \underline{w}^T \underline{x}$

- $\alpha \rightarrow \infty \implies$ more likely to be in class 1

- $\alpha \rightarrow -\infty \implies$ more likely to be in class 0.

- $\alpha = 0 \implies$ equally likely to be in class 0 or 1.

Non-Linear Trans. Use $\sigma(\underline{w}^T \phi(\underline{x}))$

ML Estimator: Given $\mathcal{D} = \{(\underline{x}_i, y_i)\}, i = 1, \dots, n$, then

$\hat{\underline{w}}_{\text{ML}} = \arg \min_{\underline{w}} - \sum_{i=1}^n \log P_{Y|\underline{X}}(y_i | \underline{x}_i, \underline{w})$

Cross Entropy b/w actual y_i and $P_{Y|\underline{X}}(\cdot | \underline{x}_i, \underline{w})$ is

$P_{Y|\underline{X}}(y_i | \underline{x}_i, \underline{w}) = \sum_{i=1}^n - (y_i \log P(1 | \underline{x}_i, \underline{w}) + (1 - y_i) \log P(0 | \underline{x}_i, \underline{w}))$

*Note: Measures the distance between 2 distributions.

*Dropped the subscripts.

Gradient Descent: No closed-form soln. so use GD.

MAP Estimator: Given $\mathcal{D} = \{(\underline{x}_i, y_i)\}, i = 1, \dots, n$, then

$\hat{\underline{w}}_{\text{MAP}} = \arg \min_{\underline{w}} - \sum_{i=1}^n \log P_{Y|\underline{X}}(y_i | \underline{x}_i, \underline{w}) + \lambda \|\underline{w}\|^2$

* $\underline{w} \sim \mathcal{N}(\underline{\mu}, \Sigma)$: Prior on \underline{w}

*Necessary: B/c same boundary $\underline{w}^T \underline{x} = 0$ for any scaling of \underline{w} .

Multiclass Logistic Regression: $Y \in \{1, 2, \dots, C\}$, then use

$$\text{softmax fn } P_Y(k \mid \underline{x}, \underline{w}_1, \dots, \underline{w}_C) = \frac{e^{\underline{w}_k^T \underline{x}}}{\sum_{c=1}^C e^{\underline{w}_c^T \underline{x}}}$$

* $W = [\underline{w}_1, \dots, \underline{w}_C] \in \mathbb{R}^{D \times C}$: Weights matrix
ML Estimator: Given $\mathcal{D} = \{(\underline{x}_i, y_i)\}, i = 1, \dots, n$, then
 $\hat{W}_{ML} = \arg \min_W - \sum_{i=1}^n \log P(y_i \mid \underline{x}_i, W)$
MAP Estimator: Given $\mathcal{D} = \{(\underline{x}_i, y_i)\}, i = 1, \dots, n$, then
 $\hat{W}_{MAP} = \arg \min_W - \sum_{i=1}^n \log P(y_i \mid \underline{x}_i, W) + \sum_{c=1}^C \lambda_c ||\underline{w}_c||^2$
Markov:

Notation:
 * $P[X_n = x_n, \dots, X_0 = x_0] = P(x_n, \dots, x_0)$
 *Index the possible values of X_n w/ integers $0, 1, 2, \dots$
Markov Chain (Memoryless/Markovian Property): A sequence of discrete-valued RVs X_0, X_1, \dots is a (discrete-time) Markov chain if

$$P[\underbrace{X_{k+1} = x_{k+1}}_{\text{Future}} \mid \underbrace{X_k = x_k}_{\text{Present}}, \underbrace{X_{k-1} = x_{k-1}, \dots, X_0 = x_0}_{\text{Past}}] =$$

$P[X_{k+1} = x_{k+1} \mid X_k = x_k] \forall k, x_1, \dots, x_{k+1}$
***Markovian:** $P(x_n, \dots, x_0) = P(x_n \mid x_{n-1}) \cdots P(x_1 \mid x_0)P(x_0)$
***Equiv. Form:** $k+1 \rightarrow n_{k+1}, k \rightarrow n_k$ and so on
 for any $n_{k+1} > n_k > \dots > n_0$ (i.e. farther in future/past)
State Distribution: State distribution of the MC at time n is
 $P_j(n) \equiv P[X_n = j], j = 0, 1, \dots \mid \underline{P}(n) \equiv [P_0(n), P_1(n), \dots]$
 *Subscript: Value of X_n , Argument: Time step
 *Row vector NOT col vector.

Transition Probabilities:
 $P_{ij}(n, n+1) \equiv P[X_{n+1} = j \mid X_n = i] \forall i, j, n$
Homogeneous MC: $P_{ij}(n, n+1) = P_{ij} \forall i, j, n$
 *Time invariant, P_{ij} does not depend on n

Transition Probability Matrix:
$$P = \begin{bmatrix} P_{00} & P_{01} & \cdots \\ P_{10} & P_{11} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Notes: (1) **Stochastic Matrix:** (1) All entries of P are non-negative and (2) each row sums to 1: $\sum_j P_{ij} = 1 \forall i$
 (2) State Dist. at time $n+1$: $\underline{P}(n) = \underline{P}(n-1)P$
 * $\underline{P}(n) = \underline{P}(0)P^n$ in terms of initial distribution $\underline{P}(0)$
 (3) State Dist. at time $n+m$: $\underline{P}(n+m) = \underline{P}(n)P^m \forall n, m$
n-step Transition Probabilities: Stochastic matrix P^n s.t.
 $P_{ij}^{(n)} \equiv P[X_{k+n} = j \mid X_k = i]$ for $n \geq 0$ are the entries of P^n

Limiting Distribution A MC has a limiting distribution \underline{q} if for any initial distribution $\underline{P}(0)$
 $\underline{P}(\infty) \equiv \lim_{n \rightarrow \infty} \underline{P}(n) = \underline{q}$ or
 $\underline{P}(0)P^\infty \equiv \underline{P}(0) \lim_{n \rightarrow \infty} P^n = \underline{q}$

Theorem: A MC has a limiting distribution \underline{q} iff

$q_i = \lim_{n \rightarrow \infty} P_{ij}^{(n)} \forall i, j$
 *i.e. every row of P^∞ equals \underline{q} (row vector)
Steady State (Stationary) Distribution $\underline{\pi}$ is a steady state distribution of a MC if $\underline{\pi} = \underline{\pi}P$
 * $1 = \sum_j \pi_j$

Theorem: If a limiting dist. exists $\underline{q} = \underline{P}(\infty)$, then it is also a steady state dist.
Ergodic: For a finite-state, irreducible, and aperiodic MC, then
 (1) Limiting dist. $\underline{q} = \lim_{n \rightarrow \infty} \underline{P}(n)$ exists and
 $q_j = \lim_{n \rightarrow \infty} P_{ij}^{(n)} \forall i, j$
 (2) Steady state dist. $\underline{\pi}$ is unique.
 (3) $\underline{\pi} = \underline{q}$

How Fast Does $\underline{P}(n)$ Converge to $\underline{\pi}$? (1) $\underline{\pi}^T = \underline{\pi}^T P^T$

* $\underline{\pi}^T$ is an eigenvector of P^T w/ eigenvalue 1
 (2) Suppose P^T has eigenvectors $U \equiv [\underline{\pi}^T, \underline{u}_2, \dots, \underline{u}_D]$ and eigenvalues $\Lambda \equiv \text{diag}[1, \lambda_2, \dots, \lambda_D]$, then
 $P^T U = U \Lambda \implies P^T = U \Lambda U^{-1}$ so n times
 $P^n = (P^T)^n = (U \Lambda U^{-1})^n = U \Lambda^n U^{-1}$
 Therefore, $\Lambda^n = \text{diag}[1, \lambda_2^n, \dots, \lambda_D^n]$

(3) For ergodic MC, $P^n \rightarrow [\underline{\pi}, \dots, \underline{\pi}]^T$ (i.e. rank 1)
 Therefore, # of non-zero eigenvalues is 1, so the rest of the eigenvalues must be $|\lambda_i| < 1 \forall i \geq 2$ s.t. $\Lambda^n = \text{diag}[1, 0, \dots, 0]$
Rate of Convergence: Depends on the 2nd largest eigenvalue of P^T i.e. $(\lambda_2)^n$ is the rate of convergence.

Bayesian Network: Network of RVs X_1, \dots, X_n w/ directed edges

***Not State-Transition Diagram:** 1 RV w/ different values w/ different probabilities to each value.
***Fully Connected Graph:** No special dependency structure, so doesn't give additional info as we can write joint dist. from defn. (true for any graph).
***Non-Fully Connected Graph (Absence of Links):** Conveys useful info about the dependency structure.
***Purpose:** Clarify the dependencies among a set of RVs to simplify the calculation of joint probabilities.
Factorization of Joint Dist. Suppose the dependencies among RVs can be represented by a DAG, then
 $P(x_1, \dots, x_n) = \prod_{i=1}^N P(x_i \mid \text{pa}\{X_i\})$
Topological Ordering: Often index the RVs s.t. each child has an index greater than those of the parents.
Fact: Every DAG has at least one topological ordering.
Conditional Independence: $A \perp B \mid C$ if
 (1) $P(a, b \mid c) = P(a \mid c)P(b \mid c) \forall a, b, c$ (i.e. A and B are indep. given C)
 (2) $P(a \mid b, c) = P(a \mid c) \forall a, b, c$ (i.e. B gives no add. info about A given C)
Common Cause (Tail to Tail):
Causal Chain (Head to Tail or Tail to Head):
Common Effect (Head to Head):