1 Probablistic Machine Learning

JHU ECE EN 520.651 2021 Fall
M. Zhou / Simplified Course Notes / Cheatsheet

1.1 Probability Review

Probablistic Model. Ω: sample space. For instance, single roll of dice $\Omega = \{1,2,3,4,5,6\}$; A: event, a set of outcomes. Venn diagrams can be used to visualize basic set operations; $P(A)$: probability measure, non-negative; For a event, the boundary case is Ω (certain set) and ϕ (null set).

Sigma Field. Given a sample space Ω and events E, F, the colection of subsets of Ω, defined as M, forms a field if (1) $\phi \in M, \Omega \in M$; (2) If $E, F \in M$, then $E \cup F \in M$ and $E \cap F \in M$; (3) If $E \in M$, then $E^C \in M$. A sigma field is closed under a countable number of unions, intersections, and complements. We care about fields and sigma fields because of consistency. When $P(E)$ and $P(F)$ are defined, but not $P(E \cap F)$, then the rule will be broken.

Strategy for the smallest σ-field. (1) Include ϕ and Ω; (2) Include disjoint parts; (3) Include all pairs, triplets, quadruplets, etc. Alternative strategy: use a binary mask, and the size of the set would be 2^N.

Probability. Set function $P : E \in \mathcal{F} \mapsto P(E) \in \mathbb{R}^+$. (1) $P(E) \geq 0$ (2) $P(\Omega) = 1$ i.e., normalization (3) $P(E \cup F) = P(E) + P(F), \forall E, F \in \Omega, s.t. E \cap F = \phi$. Then we can establish (4) $P(\phi) = 0$ (5) $P(E) = 1 - P(E^C)$ (6) $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ (7) $P(\cap_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$.

Relationship Between Events: Joint probability $P(A \cap B)$; Conditional probability $P(B|A) = P(A \cap B)/P(A)$ which means $P(B|A)P(A) = P(A|B)P(B)$. Two events $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are independent if $P(A \cap B) = P(A)P(B)$ for $P(A) > 0$ and $P(B) > 0$. Three events A, B, C (nonzero probability) are jointly independent if (1) $P(A \cap B) = P(A)P(B)$ – all pairwise combinations alike; (2) $P(A \cap B \cap C) = P(A)P(B)P(C)$ – mutual independence. Independent Experiments: the outcome of one experiment is not affected by the past, present, or future values of the other experiment.

Total Probability: Let $A_i (i = 1, \ldots, n)$ be mutually exclusive and exhaustive w/ nonzero probability, for all $B \in \Omega$, we can write

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Bayes Rule: Given the above setup and $P(B) > 0$, then

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^n P(A_i)P(B|A_i)}$$

where $P(A_i)$ is "prior", $P(B|A_i)$ is "likelihood", and $P(A_i|B)$ is "posterior".

1.2 Random Variables

Random variable $X(\cdot)$ is a function that maps outcome $\omega \in \Omega$ onto the real number line $X(\omega) \in \mathbb{R}$, i.e., $X : \Omega \mapsto \mathbb{R}$.

Cumulative Distribution Function (CDF):

$$F_X(x) = P(\omega \in \Omega | X(\omega) \leq x) \triangleq P(X \leq x)$$

(1) $F_X(-\infty) = P(\phi) = 0$, $F_X(+\infty) = P(\phi) = 1$
(2) $x_1 \leq x_2$, $F_X(x_1) < F_X(x_2)$
(3) $F_X(x) = \lim_{\epsilon \to 0^+} F_X(x + \epsilon)$
Implication: $P(a < X \leq b) = F_X(b) - F_X(a)$

Probability Density Function (PDF):

$$f_X(x) = \frac{d}{dx} F_X(x)$$

(1) $f_X(x) \geq 0$
(2) $\int_{-\infty}^{\infty} f_X(\xi) d\xi = F_X(\infty) - F_X(-\infty) = 1$
(3) $F_X(x) = \int_{-\infty}^{x} f_X(\xi) d\xi = P(X \leq x)$
(4) $\int_{x_1}^{x_2} f_X(\xi) d\xi = F_X(x_2) - F_X(x_1) = P(x_1 < X \leq x_2)$
See appendix for a list of useful distributions.

Mixed Random Variables.

$$P(X = x) = \lim_{e \to 0^+} \int_{x}^{x+e} f_X(\xi) d\xi$$

The value is generally 0 for continuous RV. When we need nonzero mass at $x = x_0$, we can use a Dirac delta $\delta(x - x_0)$.

Conditional Distribution of X given B is,

$$F_X(x|B) = \frac{P(X \leq x, B)}{P(B)} \qquad f_X(x|B) = \frac{d}{dx} F_X(x|B)$$

Joint Distribution of $X : \Omega \mapsto \mathbb{R}$ and $Y : \Omega \mapsto \mathbb{R}$

$$F_{XY}(x, y) = P(\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y) \triangleq P(X \leq x, Y \leq y)$$

where $F_{XY}(\infty, \infty) = 1$, $F_{XY}(-\infty, -\infty) = 0$. $F(x, +\infty) = F_X(x)$. $\forall x_1 \leq x_2, y_1 \leq y_2, F(x_1, y_1) \leq F(x_2, y_2)$. PDF is

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)$$

and marginal is $f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$. The conditional is $f_{X|Y}(x|y) = \frac{f_{XY}}{f_Y} = \frac{f_X f_{Y|X}}{\int_{-\infty}^{\infty} f_X(x') f_{Y|X}(y|x') dx'}$ (a slice).

Independence. X and Y are independent if $F_{XY} = F_X F_Y$ or $f_{XY} = f_X f_Y$ or $f_{X|Y} = f_X$.

1.3 Functions of Random Variables

Discrete RV. $x \in \mathcal{X}$. $P_X(x) = P(X = x)$ s.t. $\sum_{x \in \mathcal{X}} P_X(x) = 1$. Then for $Y = g(X)$ we have $P_Y(y) = \sum_{x : g(x) = y} P_X(x)$.

Continuous RV. Method I: direct computation. First compute the CDF of Y, then differentiate to obtain PDF. $P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$.

Method II. Root Formula. x_i are the roots of $y = g(x)$.

$$f_Y(y) \approx \sum_{i=1}^{N} f_X(x_i) \Big| \frac{dx_i}{d_y} \Big| \to \sum_{i=1}^{N} f_X(x_i) \frac{1}{|g'(x_i)|}$$

Multivariate Function. We cannot derive a root formula for $Z = g(X,Y)$. We use direct computation: $F_Z(z) = P(Z \le z) = P(g(X,Y) \le z) = \iint_{C_Z} f_{XY}(x,y)dxdy$. Then $f_Z(z) = \frac{d}{dz}F_Z(z)$.

Sum of Independent RVs. Special case of multivar. X and Y are independent. $Z = X+Y$. Convolution for PDF:

$$f_Z(z) = \int_{-\infty}^{+\infty} f_Y(y) f_X(z-y)dy = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-y)dx$$

1.4 Moment Generating Functions

Expectation. 1st order raw moment. For discrete RV $E[X] = \sum_{-\infty}^{+\infty} x \cdot P_X(x)$. For continuous RV $E[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx$. For $Y = g(X)$, $E[Y] = \int_{-\infty}^{+\infty} g(x) f_X dx$. For $Z = g(X,Y)$, $E[Z] = \iint_{-\infty}^{+\infty} g(x,y) f_{XY} dxdy$. Properties: (1) $E[X+Y] = E[X]+E[Y]$; (2) $E[XY] = E[X]E[Y]$ for independent X and Y.

Conditional Expect. $E[Y|X=x] = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)dy$. Law of iterated expectations $E[Y] = E_X[E_{Y|X}[Y|X]]$.

High Order Moments. The r^{th} moment of X is $m_r = E[X^r] = \int_{-\infty}^{+\infty} x^r f_X(x)dx$. The r^{th} central moment is $c_r = E[(x-\mu_x)^r] = \int_{-\infty}^{+\infty} (x-\mu_x)^r f_X(x)dx$. The variance is

$$c_2 \triangleq \sigma_X^2 = E[(x-\mu_x)^2] = E[X^2] - \mu_X^2$$

Moments do not always exist. For example, Cauthy distribution does not have that. Some properties $Var[X+a] = Var[X]$, $Var[aX] = a^2 Var[X]$, $Var[aX+bY] = a^2 Var[X] + b^2 Var[Y] + 2abCov[X,Y]$.

Joint Moments. The $(i,j)^{th}$ joint and joint central moments are $m_{ij} = E[X^i Y^j]$ and $c_{ij} = E[(X-\mu_X)^i (Y-\mu_Y)^j]$. The covariance is

$$\sigma_{XY} = Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X \mu_Y$$

Uncorrelatedness is weaker than independence.

Moment Generating Functions. $M_X(S) = E[e^{SX}]$.

$$\frac{d^r}{dS^r} M_X(S)\Big|_{S=0} = \int_{-\infty}^{+\infty} x^r f_X(x)dx \triangleq E[X^r] = m_r$$

Weak Law of Large Numbers. Let $X_1, \ldots, X_N \sim i.i.d$ with mean μ_X and $\sigma_X^2 < \infty$. The mean estimator $\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} X_i$ satisfies $P(|\hat{\mu}_X - \mu_X| \ge \delta) \le \frac{\sigma_X^2}{N\delta^2}$.

Central Limit Theorem. Characteristic function: $\Phi_X(\omega) = \int_{-\infty}^{+\infty} e^{j\omega x} f_X(x)dx$ is Fourier equivalent of MGF. Let $X_1, \ldots, X_n \sim i.i.d$ with $\mu_X = 0$, $Var(X) = 1$. Define $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$, then $\lim_{n\to\infty} \Phi_{Z_n}(\omega) = \exp\{-\frac{1}{2}\omega^2\}$.

1.5 Random Vectors

$$P(\underline{X} \le \underline{x}) = P(X_1 \le x_1, \ldots, X_N \le x_N) \quad f_{\underline{X}}(\underline{x}) = \frac{\partial^N F_{\underline{X}}(\underline{x})}{\partial x_1 \ldots \partial x_N}$$

Function of Rand Vectors. $y_i = g_i(x_1, \ldots, x_n)$, $x_i = \phi_i(y_1, \ldots)$. Root formula $f_{\underline{Y}}(\underline{y}) = \sum_{r=1}^{R} f_{\underline{X}}(\underline{x}^r)\frac{1}{|J_r|}$, $s.t.\ \underline{y} = \underline{g}(\underline{x}^r)$.

$$\frac{\partial x}{\partial y} = \begin{vmatrix} \frac{\partial \phi_1}{\partial y_1} & \cdots & \frac{\partial \phi_n}{\partial y_1} \\ \vdots & & \vdots \\ \frac{\partial \phi_1}{\partial y_n} & \cdots & \frac{\partial \phi_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_n}{\partial y_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial y_n} & \cdots & \frac{\partial g_n}{\partial y_n} \end{vmatrix}^{-1} = |J|^{-1}$$

$\underline{X} = [X_1, \ldots, X_N]^T$, Mean $E[\underline{X}] = [\mu_1, \ldots, \mu_N]^N = \underline{\mu}_X$. $\mu_i = E[X_i] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i f_{\underline{X}} dx_1 \cdots dx_N = \int_{-\infty}^{+\infty} x_i f_{X_i} dx_i$. Covariance $\mathbb{K} = E[(\underline{X}-\mu_X)(\underline{X}-\mu_X)^T] \in \mathbb{R}^N$. The correlation matrix $\boldsymbol{R} = E[XX^T] = \mathbb{K} + \mu_X \mu_X^T$.

Covariance Matrix. $\mathbb{K}$ is positive semi-definite, i.e., $z^T \mathbb{K} z \ge 0$. $\mathbb{K}$ is symmetric. $\mathbb{K} = U\Lambda U^T$ where $U^T U = I$ and any eivenvalue$\lambda_i \ge 0, \forall i = 1, \ldots, N$.

Whitening Transform. $\underline{X} \sim f_{\underline{X}}$ with $\mu_x = 0$ and PD covariance $\mathbb{K}_X = U\Lambda U^T$. Find a linear transformation $\underline{y} = C\underline{x}$ so that $\mathbb{K}_Y = I$. Thus, $\mathbb{K}_Y = E[YY^T] = E[CXX^T C^T] = C\mathbb{K}_X C^T = I$, and hence $C = \Lambda^{-\frac{1}{2}} U^T$.

Multivariate Gaussian.

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\mathbb{K}_X|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}-\mu_x)^T \mathbb{K}_X^{-1}(\underline{x}-\mu_X)\}$$

$y = Ax$ is also gaussian w/ $\mu_Y = A\mu_x$ and $\mathbb{K}_Y = A\mathbb{K}_X A^T$.

1.6 Bayesian Hypothesis Testing

Bayesian Hypothesis Testing. Observation vector $\underline{y}$, unknown state of the world H, prior $P_H(H_m)$, likelihood $P_{\underline{Y}|H}(\underline{y}, H_m)$, posterior $P_{H|Y}(H_m|y)$. From Bayes rule,

$$P_{H|Y}(H_m|y) = \frac{P_H(H_m) P_{Y|H}(y|H_m)}{\sum_{m'} P_H(H_{m'}) P_{Y|H}(y|H_{m'})}$$

Binary Hypothesis Testing. $H \in \{H_0, H_1\}$. $P_H(H_0) \triangleq P_0$, $H_0 : P_{Y|H}(y|H_0)$. A decision rule $\hat{H}(\cdot)$ maps observation $\underline{y}$ onto a hypothesis $H \in \{H_0, H_1\}$. Cost function $c_{ij} \triangleq \tilde{C}(H_j, H_i)$ where H_j true state, H_i our guess. A cost function is valid when $c_{01} > c_{11}$ and $c_{10} > c_{00}$.

Likelihood Ratio Test (LRT). Minimize $\phi(\hat{H}) = E_{Y,H}[\tilde{C}(H, \hat{H}(y))] = E_Y[E_{H|Y}[\tilde{C}(H, \hat{H}(y))|Y = y]] = \sum_y p_Y(y) E_{H|Y}[\tilde{C}(H, \hat{H}(y))|Y = y]$.

$$L(y) \triangleq \frac{P_{Y|H}(y|H_1)}{P_{Y|H}(y|H_0)} \gtrless_{(H_0)}^{(H_1)} \frac{P_0(c_{10}-c_{00})}{P_1(c_{01}-c_{11})} \triangleq \eta$$

MAP decision rule: symmetric cost $c_{00} = c_{11} = 0$, $c_{01} = c_{10} = 1$. $P_1 P(y|H_1) \gtrless_{(H_0)}^{(H_1)} P_0 P(y|H_0)$, i.e., $P(H_1|y) \gtrless_{(H_0)}^{(H_1)} P(H_0|y)$. ML decision rule: symmetric cost and $P_0 = P_1 = \frac{1}{2}$. $P(y|H_1) \gtrless_{(H_0)}^{(H_1)} P(y|H_0)$.

1.7 NonBayesian Hypothe sis Testing

Classical Hypothesis Testing. $L(y) \triangleq \frac{P(y|H_1)}{P(y|H_0)} \gtrless_{(H_0)}^{(H_1)} \eta$. This is the generalization of the Bayesian case.

Operating Characteristics. $P_D = P(\hat{H}(y) = H_1|H_1) = \int_{y_1} P(y|H_1)dy$. $P_F = P(\hat{H}(y) = H_1|H_0) = \int_{y_1} P(y|H_0)dy$.

ROC. P_F-P_D curve. Axes limits $[0, 1]$. $\eta = 0$ (upper right corner), $\eta \to \infty$ (lower left corner). Mototonically non-decreasing in η. Lies above the diagonal.

Neyman-Pearson HT. $\max_{\hat{H}(\cdot)} P_D$ $s.t.$ $P_F \le \alpha$. The optimal solution can be expressed as LRT, where $\eta = \lambda$ such that $P_F = \alpha$. It is intersection of ROC w/ $P_F = \alpha$ line.

Randomized Decision Rule for discrete-valued data. $\eta_{i+1} > \eta_i$. $P_D = p \cdot P_D(\eta_i) + (1-p)P_D(\eta_{i+1})$. $P_F = p \cdot P_F(\eta_i) + (1-p)P_F(\eta_{i+1})$.

Efficient Frontier for LRT. The achievable (P_F, P_D) operating points. On efficient frontier, (1) the $(0, 0)$ and $(1, 1)$ points always lie here; (2) $P_D \ge P_F$; (3) is concave (randomized decision should not beat NP); (4) $\frac{dP_D}{dP_F} = \eta$.

1.8 Minmax Hypothesis Testing

Setup. Adversarial game w/ cost but w/o priors

$$\hat{H}_M(\cdot) = \arg\min_{f(\cdot)}[\max_{p=Pr(H=H_1)\in[0,1]} E_{Y,H}[\tilde{C}(H, f(Y))]]$$

The class of $f(\cdot)$ is restricted to LRTs, $\hat{H}_B(y, q) = H_1 \cdot \mathbf{1}\{\mathcal{L}(y) > \frac{1-q}{q}\frac{c_{10}-c_{00}}{c_{01}-c_{11}}\} + H_0\{o.w.\}$. The solution is

$$P_D(q^*) = \frac{c_{10} - c_{00}}{c_{01} - c_{11}} - P_F(q^*)[\frac{c_{10} - c_{00}}{c_{01} - c_{11}}]$$

1.9 Bayesian Parameter Estimation

Problem Setting. Hidden parameter $X \in \mathcal{X}$ continuous (RV); Noisy observation $Y \in \mathcal{Y}$ either continuous or discrete; Prior belief $P_X(x)$; Likelihood (observation model) $P_{Y|X}(y|x)$; The goal is to construct an estimator $\hat{x}(\cdot)$ that produces an estimate of x given the observation $Y = y$. Similar to Bayesian HT, an objective criterion $C(a, \hat{a})$ is required to build and evaluate this estimator, so

$$\begin{aligned}\hat{x}(\cdot) &= \arg\min_{f(\cdot)} E_{XY}[\tilde{C}(x, f(y))] \\ &= \arg\min_a \int_{-\infty}^{\infty} C(x, a)P_{X|Y}(x|y)dx \\ &= \arg\min_a E_{X|Y}[C(x, a)|Y = y]\end{aligned}$$

Minimum Absolute Error (MAE), $C(a, \hat{a}) = |a - \hat{a}|$

$$\hat{x}_{MAE}(y) = \arg\min_a \int_{-\infty}^{\infty} |x - a|P_{X|Y}(x|y)dx$$
$$\Rightarrow \int_{-\infty}^{a} P_{X|Y}(x|y)dx - \int_{a}^{\infty} P_{X|Y}(x|y)dx = 0$$

$\hat{x}_{MAE}(y)$ is the MEDIAN of $P_{X|Y}$ – not always unique.

Minimum Uniform Cost, $C(a, \hat{a}) = \mathbf{1}\{|a - \hat{a}| > \epsilon\}$

$$\hat{x}_{MUC}(y) = \arg\max_a \int_{a-\epsilon}^{a+\epsilon} P_{X|Y}(x|y)dx$$

The $\hat{x}_{MUC}(y)$ is the center of the 2ϵ interval with the most mass of $P_{X|Y}$. When ϵ tends to zero, MUC is defined as the MAP estimator, $\lim_{\epsilon\to 0}\hat{x}_{MUC}(y) = \arg\max_a P_{X|Y}(x|y) \triangleq \hat{x}_{MAP}(y)$.

Bayes Least Squares (BLS):

$$C(\boldsymbol{a}, \hat{\boldsymbol{a}}) = \|\boldsymbol{a} - \hat{\boldsymbol{a}}\| = (a - \hat{a})^T(a - \hat{a})$$

We plug the cost into the optimization problem, then derive the internal part and letting the derivative be zero, then

$$\hat{\boldsymbol{x}}_{BLS}(\boldsymbol{y}) = \int_{-\infty}^{\infty} xP_{X|Y}dx = E[\boldsymbol{X}|\boldsymbol{y}]$$

Performance Characeteristics. Vector by default. <u>Error.</u> Given an instance x, $e(x, y) = \hat{x}(y) - x$. <u>Bias</u> $b = E_{XY}[e(x, y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e(x, y)P_{x,y}dxdy$. <u>Variance</u> $\Lambda_e = E_{xy}\{[e(x, y) - b][e(x, y) - b]^T\}$. $MSE = E_{xy}[e(x, y)e(x, y)^T] = \Lambda_e + bb^T$. (1) $\hat{x}_{BLS}(y)$ is unbiased, $b = 0$. (2) $\hat{x} = \hat{x}_{BLS}$ iff $e(x, y)$ is orthogobal to any function of y, $E_{XY}[(\hat{x}(y) - x)g^T(y)] = 0$.

1.10 Linear Least Squares Estimation

The previous estimators require complete characterization of P_X and $P_{Y|X}$, and posterior is difficult to compute.

Linear Least-Squares Estimation.

$$\hat{x}_{LLS} = \arg\min_{f(\cdot)} E_{XY}[\|x - f(y)\|^2] \; s.t. \; f(y) = \mathbb{A}\underline{y} + \underline{d}$$

A linear estimator $\hat{x}(y)$ is LLS iff unbiased and orthogonal to data:

$$E_{XY}[\hat{x}(y) - x] = 0 \quad E_{XY}[(\hat{x}(y) - x)y^T] = 0_{N\times N}$$

Constructing LLS. Use defintion if moment unavailable.

$$\hat{x}_{LLS}(\underline{y}) = \underline{\mu}_X + \Lambda_{XY}\Lambda_Y^{-1}(\underline{y} - \underline{\mu}_Y)$$

Error covariance $\Lambda_{LLS} = \Lambda_X - \Lambda_{XY}\Lambda_Y^{-1}\Lambda_{XY}^T$. When X and Y are jointly Gaussian, $\hat{x}_{LLS} = \hat{x}_{BLS}$.

1.11 NonBayesian Parameter Estimation

Y is parameterized by x, $P_{Y|X}(y|x) \to P_Y(y; x)$. An estimator is valid if it does not depend explicitly on the parameter we are trying to estimate. We want to minimize MSE, i.e., $tr(e(y)e(y)^T) = \Lambda_e(x) + b_{\hat{x}}b_{\hat{x}}^T$. Estimator $\hat{x}(y)$ is unbiased if $b_x(x) = 0$. $\Lambda_e(x) = \Lambda_{\hat{x}}$.

Minimum Variance Unbiased Estimators.

$$\hat{x}_{MVU}(y) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}(x) = \arg\min_{\hat{x}\in\mathcal{A}} E_Y[e^2(y)] \; \forall x$$

where admissible estimators $\mathcal{A} = \{\hat{x}(\cdot)\text{: valid and unbiased}\}$. It might not exist.

Cramer-Rao Bound (CRB). Score function $S(y;x) = \frac{\partial}{\partial x}\ln P_Y(y;x)$. Fisher Information $J_Y(x) = E_Y[S^2(y;x)] = -E_Y[\frac{\partial^2}{\partial x^2}\ln P_Y(y;x)]$. If $\hat{x}(\cdot)$ is valid and unbiased, then $\lambda_{\hat{x}} \geq \frac{1}{J_Y(x)}\ \forall x$.

Efficient Estimators. $\hat{x}(y) = x + \frac{1}{J_Y(x)}\frac{\partial}{\partial x}\ln P_Y(y;x)$. (1) $\hat{x}_{eff}(y)$ is guaranteed to be unbiased. (2) if $\hat{x}_{eff}$ exists and is valid, then it is unique. (3) if $\hat{x}_{eff}(y)$ exists, then it meets CRB and is MVU.

1.12 Maximum Likelihood Estimation

Proxy for efficient estimator with nice properties:

$$\hat{x}_{ML}(y) = \arg\max_{x\in\mathcal{X}} P_Y(y;x)$$

ML estimator is simpler to compute or numerically approximate. Meanwhile, it is intuitive because picking x that gives you the largest chance of observing data $Y = y$. If $\hat{x}_{eff}(y)$ exists, then it equals $\hat{x}_{ML}(y)$.

Invertible Mapping. $\theta = g(x) \rightarrow \hat{\theta}_{ML}(y) = g(\hat{x}_{ML}(y))$.

Vector Parameters. Score function $\vec{S}_{\vec{Y}}(\vec{x}) = [\frac{\partial \log P_Y(\vec{y};x)}{\partial x_1},\ldots,\frac{\partial \log P_Y(\vec{y};x)}{\partial x_N}]$. Fisher information $J_{\vec{Y}}(\vec{x}) = E_Y[\vec{S}_Y(\vec{x})^T\vec{S}_Y(\vec{x})]$. CRB: Covariance matrix $\Lambda_{\hat{x}}(x)$ of any unbiased estimator satisfies the following inequality: $\Lambda_{\hat{x}(x)} \geq J_Y^{-1}(x) \leftrightarrow (\Lambda_{\hat{x}(x)} - J_Y^{-1}(x))$ is PSD. (1) $J_Y(x) = -E_Y[\frac{\partial^2}{\partial x^2}\ln P_Y(y;x)]$; (2) $\hat{x}_{eff}(y)$ exists if it can be expressed $\hat{x}_{eff}(y) = x + J_Y^{-1}(x)[\frac{\partial}{\partial x}\ln P_Y(y;x)]^T$; (3) if $\hat{x}_{eff}(y)$ exists, then it is the ML estimator.

Misc. MAP estimator maximizes the joint distribution.

1.13 Exponential Families

A parameterized family of distributions $\{Pr(\cdot;x), x\in\mathcal{X}\}$ over the alphabet $\mathcal{Y}$ is a one-parameter exponential family if it can be written as

$$P_Y(y;x) = \exp\{\lambda(x)t(y) - \alpha(x) + \beta(y)\}\ \forall x\in\mathcal{X}, y\in\mathcal{Y}$$

Construction. I. Geometric mean, $p_1(y)$ and $p_2(y)$ strictly positive on $\mathcal{Y}$. $P_Y(y;x) = \frac{1}{Z(x)}p_1(y)^x p_2(y)^{1-x}$, $x\in\mathcal{X} = [0,1]$. So $\log P_Y(y;x) = x\log(\frac{p_1(y)}{p_2(y)}) + \log p_2(y) - \log Z(x)$.

II. Tilting based on distribution $q(y)$. $P_Y(y;x) = q(y)e^{xy}/Z(x)$. $\log P_Y(y;x) = xy + \log q(y) - \log Z(x)$.

Moment Generation. A linear exponential family has $\lambda(x) = x$. (1) $\alpha'(x) = E_Y[t(Y)]$; (2) $\alpha''(x) = Var[t(Y)]$; (3) $J_Y(x) = Var[t(Y)]$.

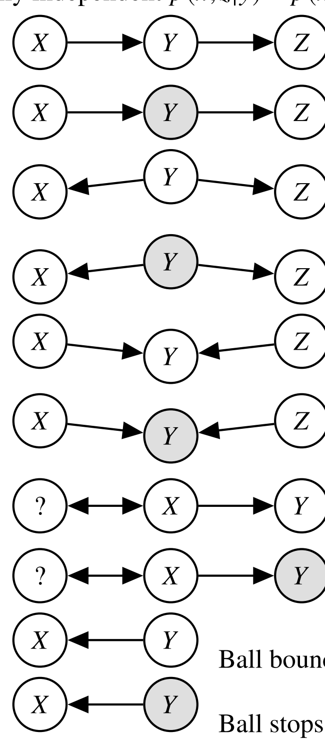
Efficient Estimators. If $\hat{x}_{eff}$ exists, then $P_Y(y;x)$ is a member of an exponential family with $\lambda(x) = \int^x J_Y(u)du$ and $t(y) = \hat{x}_{ML}(y) = \hat{x}_{eff}(y)$.

Conjugate Priors. Let $Q = \{q(\cdot;\theta) : \theta\in\Theta\subset\mathbb{R}^K\}$ be a family of distributions based on K parameters $\theta = [\theta_1,\ldots,\theta_K]^T$ such that $q(x;\theta)$ is continuously invertible in θ. Then Q is a conjugate prior family for $P_{Y|X}$ if $P_X(\cdot)\in Q \rightarrow P_{X|Y}(\cdot|y)\in Q$.

1.14 Directed Graphical Models

Defines a family of joint probability distributions over set of RVs. The setup if Directed Acyclic Graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ with nodes (RVs) and edges. For each node $i\in\mathcal{V}$, let π_i be the set of parent nodes. Then the family of distributions satisfy $P_X(x_1,\ldots,x_n) = \prod_{i=1}^n P_X(x_i|x_{\pi_i})$. In this graph, absenceof edge means conditional independence.

Bayes' Ball Algorithm. It determines conditional independence given obserations. Examine whether x and z are marginally independent $p(x,z) = p(x)p(z)$, or conditionally independent $p(x,z|y) = p(x|y)p(z|y)$.



Can pass through Y.

Blocked at Y.

Can pass through Y.

Blocked at Y.

Blocked at Y.

Can pass through Y.

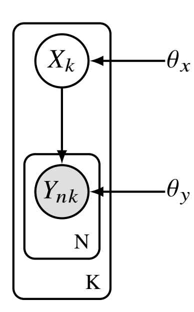
Ball stops at leaf Y.

Ball bounces back at Y.

Ball bounces back at root Y.

Ball stops at root Y.

Plate Notation. See wikipedia. Circles: random variable; Squares: non-random parameters.



$p(x_\cdot, y_\cdot)$
$= \prod_k [p_X(x_k;\theta_k)$
$\prod_n p_{Y|X}(y_{nk}|x_k;\theta_y)]$

1.15 Mixture Models

K-Means. Centroid of k-th cluster μ_k. Cluster assignment of the n-th point $Z_n \in \{1,\dots,K\}$. We define $\underline{Z}_n = [Z_n^1,\dots,Z_n^k]$ as binary indicator vector. (1) Given centroids, we can assign clusters as $Z_n^k = \mathbf{1}\{k = \arg\min_{k'} \|y_n - \mu_{k'}\|\}$. (2) Given assignments, we compute the centroids as $\mu_k = (\sum_n Z_n^k y_n)/\sum_n Z_n^k$. In this algorithm, we are actually doing coordinate descent to minimize the L2 objective $J = \sum_n \sum_k Z_n^k \|y_n - \mu_k\|^2$. This leads to solution to a Gaussian mixture model w/ equal variances.

Mixture Model. Observe i.i.d. scalars $Y_1,\dots,Y_N$. Then $P_Y(Y_1,\dots,Y_N|\theta) = \prod_n \sum_k \pi_k p_k(y_n;\theta_k)$. Goal is to find ML parameter estimates for π_k and θ_k. The log likelihood is $\ell(\underline{Y},\theta) = \sum_n \log(\sum_k \pi_k p_k(y_n;\theta_k))$. This is not easy to optimize. We introduce auxilliary one-hot RV Z_n,

$$P(Y_1,\dots,Y_N,Z_1,\dots,Z_N;\theta) = \prod_n \prod_k [\pi_k p_k(y_n;\theta_k)]^{Z_n^k}$$

and the corresponding $\ell_c(\underline{Y},\underline{Z};\theta)$ is easier to optimize as the summation inside log has been eliminated. Then we want to maximize the lower bound of $\ell_c(\underline{y},\theta)$, namely $E_{Z|Y}[\ell_c(\underline{y},\underline{z};\theta)]$. We let the posterior $\tau_n^k = E[Z_n^k]$.

$$E_{Z|Y}[\ell_c(y,z;\theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} E[z_n^k](\log\pi_k + \log p_k(y_n;\theta_k))$$

Clustering Algorithm. Step (E): find posterior $\{\tau_n^k\}$ given $\underline{y}$ and parameters $\{\pi_*,\theta_*\}$.

$$\begin{aligned}\tau_n^k &= E_{Z|Y}[Z_n^k] = Pr(Z_n^k = 1|y_n;\theta)\\ &= \frac{Pr(Z_n = k)p_k(y_n;\theta_k)}{P_Y(y_n;\theta)} = \frac{\pi_k p_k(y_n;\theta_k)}{\sum_{k'} \pi_{k'} p_{k'}(y_n;\theta_k)}\end{aligned}$$

Step (M): Find mixing weights $\{\pi_k\}$ given $\underline{y}$, $\{\tau_n^k\}$ and $\{\theta_*\}$ using Lagrangian,

$$\phi(\underline{y},\underline{z};\theta) = E_{Z|Y}[\ell(\underline{y},\underline{z};\theta)] - \lambda(\sum_k \pi_k - 1)$$

And setting $\partial\phi/\partial\pi_k$ to zero yields $\pi_k = \frac{1}{N}\sum_n \tau_n^k$. Then we optimize $\{\theta_*\}$ based on specified densities.

Gaussian Mixture. $p_k(y_n;\theta_k) = \mathcal{N}(y_n;\mu_k,\sigma_k^2)$.

$$\mu_k = \frac{\sum_n \tau_n^k y_n}{\sum_n \tau_n^k} \qquad \sigma_k^2 = \frac{\sum_n \tau_n^k (y_n - \mu_k)^2}{\sum_n \tau_n^k}$$

1.16 Generalized Mixture Model

Setup. Observed data $y \in \mathcal{Y}$, unknown (deterministic) parameters $x \in \mathcal{X}$, latent/hidden random variables $z \in \mathcal{Z}$. Objective: $\hat{x} = \arg\max_x \log P_Y(y;x)$.

Jensen's Inequality.

$$\log(\lambda z_1 + (1-\lambda)z_2) \geqslant \lambda\log(z_1) + (1-\lambda)\log(z_2)$$

Construct Lower Bound. Introduce $q(z|y)$ an arbitrary distribution. $\ell(y;x) = \log P_Y(y;x) = \log(\sum_z P_{Y,Z}(y,z;x)) = \log(\sum_z q(z|y)\frac{P_{Y,Z}(y,z;x)}{q(z|y)}) \geqslant \sum_z q(z|y)\log(\frac{P_{Y,Z}(y,z;x)}{q(z|y)}) \triangleq \mathcal{L}(q,x)$.

EM Algorithm. Coordinate ascent on $q(\cdot|y)$ and x at $(t+1)^{st}$ iteration. E-step $q^{(t+1)} = \arg\max_{q(\cdot)} \mathcal{L}(q,x^{(t)})$. M-step $x^{(t+1)} = \arg\max_x \mathcal{L}(q^{(t+1)},x)$. Step order can be switched depending on initialization.

M-Step. $x^{(t+1)} = \arg\max_x \sum_z q^{(t+1)}(z|y)[\log P(y,z;x) - \log q^{(t+1)}(z|y)] = \arg\max_x \sum_z q^{(t+1)}(z|y)\log P(y,z;x) = \arg\max_x E_{q^{(t+1)}}[\log P(y,z|x)]$.

E-Step. $\mathcal{L}(q,x^{(t)}) = \sum_z q(z|y)\log(\frac{p_Y(y;x^{(t)})p_{Z|Y}(z|y;x^{(t)})}{q(z|y)}) = \sum_z q(z|y)\log(p_Y(y;x^{(t)})) + \sum_z q(z|y)\log(\frac{p_{Z|Y}(z|y;x^{(t)})}{q(z|y)})$. The left item $= \log P_Y(y;x^{(t)})\sum_z q(z|y) = \ell(y;x^{(t)})$. Gibb's Inequality: $E_p[\log p(z)] \geqslant E_q[\log p(z)]$ with equality iff $p(z) = q(z)$. So right item < 0 unless $q(z|y) = p(z|y;x^{(t)})$. So the solution is $q^{(t+1)}(z|y) = P_{Z|Y}(z|y;x^{(t)})$, and $\mathcal{L}(q^{(t+1)},x^{(t)}) = \log P_Y(y;x^{(t)})$.

MAP Estimation. $\hat{x}_{MAP}(y) = \arg\max_x p(x|y) = \arg\max_x \frac{p(x,y)}{p(y)} = \arg\max_x p(x,y)$.

1.17 Deterministic Approximations

The goal of "belief approximation" is to find a simpler distribution $q(\cdot)$ that is sufficiently "close" to posterior $P_{X|Y}(\cdot|y)$.

KL Divergence. Given a true distribution $p(\cdot)$, and an approximating distribution $q(\cdot)$,

$$D(p\|q) = E_p[\log\frac{p(x)}{q(x)}] = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

Gibb's inequality $\rightarrow$ $D(p\|q) = E_p[\log p(x)] - E_p[\log q(x)] \geq 0$. Only when $p = q$, $D(p\|q) = 0$. Besides, $D(p\|q) \neq D(q\|p)$. $I(X,Y) = D(P_{XY}\|P_X P_Y)$.

Laplace's Method.

$$\hat{P}_X(x) = \frac{\hat{P}_0(x)}{Z_{\hat{P}}} = P_0(x)\exp\{-\frac{1}{2}J(\hat{x})(x-\hat{x})^2\}$$

where $-J(\hat{x}) = \frac{d^2}{dx^2}\log P_0(x)|_{x=\hat{x}}$ is the observed fisher info, partition function $Z_p = P_0(x)\sqrt{2\pi(1/J(\hat{x}))}$.

$$P_{X|Y}(x|y) \approx \mathcal{N}(x;\hat{x}_{MAP}(y),\hat{\sigma}^2(y))$$

$$\hat{\sigma}^2(y) = [-\frac{\partial^2}{\partial x^2}\log P_X(x)|_{\hat{x}_{MAP}} - \frac{\partial^2}{\partial x^2}\log P_{Y|X}(y|x)|_{\hat{x}_{MAP}}]^{-1}$$

The empirical fisher info is $J_{Y=y}(\hat{x}_{MAP}(y))$. For large dataset use $\hat{x}_{ML}$ instead.

Variational Methods.

$$\hat{p}(x) = \arg\min_{q \in Q} D(q \| p_{X|Y})$$
$$\hat{p}(x) = \arg\min_{q \in Q} D(q \| p_{X,Y})$$

And variational free energy is $\mathcal{FE} = D(q||p_{XY})$.

1.18 Stochastic Approximations

Goal: approximate $E_p[f(x)]$ for general $f(\cdot)$.

Approach: suppose we had samples of $x_1, \ldots, x_n \sim iid\, p(x)$.

$$E_p[f(x)] \approx \hat{f} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$
$$E_p[\hat{f}] = \frac{1}{n}\sum_{i=1}^{n} E_p[f(x_i)] = E_p[f(x)]$$
$$Var[\hat{f}] = \frac{1}{n^2}\sum_{i=1}^{n} Var[f(x_i)] = \frac{Var[f(x)]}{n} \to_{n\to\infty} 0$$

We obtain samples from $q(\cdot)$, which is easy to handle, and transform them into samples of $p(\cdot)$.

$$q(x) > 0 \;\forall\, x \in \mathcal{X} \; s.t. \; p(x) > 0$$

Importance Sampling. We call $q(\cdot)$ as "proposal" or "sampling" distribution.

$$q(x) = \frac{q_0(x)}{Z_q} = \frac{q_0(x)}{\int_{x\in\mathcal{X}} q_0(x)dx} \qquad w(x_i) = \frac{p_0(x_i)}{q_0(x_i)}$$
$$\hat{f} \triangleq \sum_{i=1}^{n} \frac{w(x_i)}{\sum_{j=1}^{n} w(x_j)} f(x_i)$$

Rejection Sampling. Draw samples from $p(\cdot)$, with unnormalized "proposal distribution" $q_0(\cdot)$ where $cq_0(x) > p_0(x)$ and c is constant. (1) sample $x \sim q(\cdot)$. (2) sample $u \sim U[0, cq_0(x)]$. (3) keep sample if $u \le p_0(x)$, otherwise discard.

Markov-Chain Monte Carlo (MCMC). Generate samples from $p(\cdot)$, and does not require proposal distribution $q(\cdot)$ to be close to $p(\cdot)$. It produce correlated samples. Metropolis-Hastings: proposal distribution at time $(n+1)$ is $q(\cdot;x_n)$, parametrized by previous state. (1) given x_n, generate candidate x' from $q(\cdot;x_n)$. (2) compute acceptance probability $\alpha(x_n \to x') = \min\{1, \frac{p_0(x')q(x_n;x')}{p_0(x_n)q(x';x_n)}\}$. (3) transition to x' wp. α otherwise stay in x_n. State transition probability is hence $q(\cdot;x_n)\alpha(x_n;x)$.

Gibbs Sampling.

$$p(x)\alpha(x;y)q(y;x) = p(x)\min\{1, \frac{p(y)q(x;y)}{p(x)q(y;x)}\}q(y;x)$$

1.19 Hidden Markov Models

Graphical Models -> (Discrete states from ML perspective) -> HMMs; Parametrization for Homogeneous HMM does not depend on time.

Binary indicator vector and initial state probabilities

$$q_t = [q_t^1, \ldots, q_t^M]^T \qquad \pi = [\pi^1, \ldots, \pi^M]^T \qquad p(q_0^i = 1) = \pi^i$$

We augment the prior π w/ $M \times M$ transition probability matrix $\mathbf{A}$, where

$$a_{ij} = p(q_{t+1} = j | q_t = i)$$

With q as hidden state, we observe y . The likelihood is $p(y_t|q_t;\eta)$. And the joint density of HMM is

$$p(\boldsymbol{q}, \boldsymbol{y}) = p(q_0)\prod_{t=0}^{T-1} p(q_{t+1}|q_t)\prod_{t=0}^{T} p(y_t|q_t;\eta)$$

Inference. Assume parameters $\{\pi, A, \eta\}$ are known. For a given sequence $\underline{q}$, we compute $p(\underline{q}|\underline{y})$. This can be simplified to computing $p(q_t|\underline{y}) = \frac{p(q_t)p(\underline{y}|q_t)}{p(\underline{y})} = \frac{p(y_0,\ldots,y_t,q_t)p(y_{t+1},\ldots,y_T|q_t)}{p(\underline{y})}$ where $\alpha(q_t) = p(y_0, \ldots, y_t, q_t)$ and $\beta(q_t) = p(y_{t+1}, \ldots, y_T|q_t)$. Partition function $p(\underline{y}) = \sum_{q_t}\alpha(q_t)\beta(q_t)$.

Forward Recursion. $\alpha(q_{t+1}) = \sum_{q_t}\alpha(q_t)a_{q_{t+1},q_t}p(y_{t+1}|q_{t+1})$, starting from $\alpha(q_0) = p(y_0, q_0) = \pi_{q_0}p(y_0|q_0)$.

Backward Recursion. $\beta(q_t) = \sum_{q_{t+1}}\beta(q_{t+1})a_{q_{t+1},q_t}p(y_{t+1}|q_{t+1})$, starting from $\beta(q_T) = [1, \ldots, 1]^T$.

Parameter Estimation. $\hat{\theta} = \{\hat{A}, \hat{\pi}, \hat{\eta}\} = \arg\max_{A,\pi,\eta}\log\sum_{q_0}\cdots\sum_{q_T}\pi_{q_0}\prod_{t=0}^{T-1}a_{q_{t+1},q_t}\prod_{t=0}^{T}p(y_t|q_t;\eta)$. E-step. $E_{q|y;\theta^{(p)}}[\log p(q,y)]$. M-step. update parameters based on statistics.

1.20 Kalman Filtering

Graphical Models -> (Continuous states from SP perspective) -> Kalman Filter (Tracking / Online Learning).

Kalman Filtering. Vector x_t and y_t are true and measured position at time t respectively. u_t is driving input, deterministic and known.

$$x_t = A_t x_{t-1} + B_t u_t + \mathcal{E}_t \qquad \mathcal{E}_t \sim iid\; \mathcal{N}(0, Q_t)$$
$$y_t = C_t x_t + D_t u_t + \delta_t \qquad \delta_t \sim iid\; \mathcal{N}(0, R_t)$$

(1) Predict (prime): compute $p(x_t|y_{1:t-1}, u_{1:t};\theta) \sim \mathcal{N}(\mu_t', \Sigma_t')$. $\mu_t' = E[x_t] = A_t\hat{\mu}_{t-1} + B_t\mu_t$. $\Sigma_t' = E[(x_t - \mu_t')(x_t - \mu_t')^T] = A_t\hat{\Sigma}_{t-1}A_t^T + Q_t$. (2) Refine to $p(\hat{x}_t|y_{1:t};\theta)$ (hat). For Gaussians, BLS=LLS=$\hat{\mu}_t$. So $\hat{\mu}_t = \mu_t' + \Lambda_{XY}\Lambda_Y^{-1}(y_t - E[y_t])$, where $E[y_t] = C_t\mu_t' + D_t u_t$, $\Lambda_Y = C_t\Sigma_t' C_t^T + R_t$, $\Lambda_{XY} = E[(x_t - \mu_t')(y - E[y_t])^T] = \Sigma_t' C_t^T$. Let Kalman

Gain Matrix be $\mathbb{K}_t = \Sigma_t' C_t^T [C_t \Sigma_t' C_t^T + R_t]^{-1}$, and residual $r_t = y_t - C_t \hat{\mu}_{t-1} - D_t u_t$. Hence $\hat{\mu}_t = \mu_t' + \mathbb{K}_t r_t$. $\hat{\Sigma}_t = \Sigma_t' - \Lambda_{XY} \Lambda_Y^{-1} \Lambda_{XY}^T = (I - \mathbb{K}_t C_t)\Sigma_t'$.

Extended Kalman Filter. Non-linear function $g(\cdots)$ and $h(\cdots)$. $x_t = g(x_{t-1}, u_t) + \mathcal{E}_t$, $y_t = h(x_t, u_t) + \delta_t$. $\{x_t, y_t\}$ are nolonger jointly Gaussian, but we can linearly approximate. (1) Let Jacobian $G_{ij} = \frac{\partial g_i(x,u)}{\partial x_j}$ and $G_t = G|_{x=\hat{\mu_{t-1}}}$, $x_t \approx g(\hat{\mu}_{t-1}, u_t) + G_t(x_{t-1} - \hat{\mu}_{t-1}) + \mathcal{E}_t$. So $\mu_t' = g(\hat{\mu}_{t-1}, u_t)$. $\Sigma_t' = G_t \hat{\Sigma}_t G_t^T + Q_t$. (2) Let Jacobian $H_{ij} = \frac{\partial h_i(x,u)}{\partial x_j}$ and $H_t = H_{x=\mu_t'}$, $y_t \approx h(\mu_t', u_t) + H_t(x_t - \mu_t') + \delta_t$. So $\mathbb{K}_t = \Sigma_t' H_t^T (H_t \Sigma_t' H_t^T + R_t)^{-1}$, $\hat{\mu}_t = \mu_t' + \mathbb{K}_t(y_t - h(\mu_t', u_t))$, $\hat{\Sigma}_t = (I = \mathbb{K}_t H_t)\Sigma_t'$.

Conjugate Priors. (successive belief revision). Let $Q = \{q(\cdot;\theta); \theta \in \mathbb{R}^K\}$ denote a family of distributions specified by some dimension K. Q is a conjugate prior family for the above iid model if $\forall y \in \mathcal{Y}$, $p_X(\cdot) \in Q \to P_{X|Y}(\cdot|y) \in Q$. If $\mathcal{X}$ is finite, then a conjugate prior family always exists. Namely $q(\cdot)$ is categorical of dimension $|\mathcal{X}|$. Suppose data is coming sequentially,

$$p_{X|Y_1} = \frac{p_X p_{Y_1|X}}{\int_x P_X P_{Y_1|X} dx} \quad P_{X|Y_1,Y_2} = \frac{p_{X|Y_1} p_{Y_2|X,Y_1}}{\int_x p_{X|Y_1} p_{Y_2|X,Y_1} dx}$$

The X separates Y_1 and Y_2 so Y_1 can be removed from terms involving Y_2 from above right side Eq.

1.21 Dirichlet Process

Probability Simplex. Categorical $Z \in \{1, \ldots, K\}$. $P_Z = [P_Z(1), \ldots, P_Z(K)]^T$ s.t. $P_Z(k) \geq 0$ and $\sum P_Z(k) = 1$. It lies on an affine hyperplane of dimension $(K-1)$ known as probability simplex. More "uniform" distributions lie at the center, while "skewed" distribution concentrate on edges or at vertices.

Dirichlet Distribution. Continuous-valued distribution w/ support over P_K. $X = [x_1, \ldots, x_K]^T$, $\forall x_k \geq 0$, $\sum_k x_k = 1$; $\alpha = [\alpha_1, \ldots, \alpha_K]^T$, $\alpha_k \geq 0$.

$$Dir(X;\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \quad B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

We define $\alpha_0 = \sum_k \alpha_k$ which controls the strength or "concentration" of the distribution. Ratio among $\{\alpha_1, \ldots, \alpha_K\}$ controls peak location. $E[x_k] = \frac{\alpha_k}{\alpha_0}$; $Var(x_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0+1)}$.

Multinomial Conjugacy. $p_X(x;\alpha) = Dir(x;\alpha)$. $Z \in \{1, \ldots, K\}$ and $P(z = k|x) = x_k$, $p(z|x) = \prod_k (x_k)^{z_k}$. Observe iid z_N, $p(x,z) = p_X(x;\alpha)\prod_k p(z_i|x)$. So $p(x|z) = Dir(x; \{\alpha_k + \sum_{i=1}^N z_i^k\}_k)$. Observing data changes both the center and concentration of underlying parameter.

Dirichlet Process. Denote $G \sim DP(\alpha, H)$. $P(\theta_k;\lambda) \triangleq H(\theta_k)$, $X \sim Dir(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$. $G(\theta) = \sum_{k=1}^K x_k \delta_{\theta_k}(\theta)$. $G(\cdot)$ is mixture of delta funcs centered on $\{\theta_k\}$. $Pr(\bar{\theta}_i = \theta_k) = x_k$. Sampling will always result in K clusters. Dir process is Distribution over probability measures $G : \theta \mapsto R^+$ defined by $[G(T_1), \ldots, G(T_K)]^T \sim Dir(\alpha H(T_1), \ldots, \alpha H(T_K))$ for any partition $\{T_1, \ldots, T_K\}$ of parameter domain θ.

Extend Dirichlet Conjugacy. $X \sim Dir(\alpha_1, \ldots, \alpha_K)$, $Z_i \sim Multi(x)$. $x|z_1 \ldots z_n \sim Dir(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$. And $G|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H \sim DP(\alpha + N, \frac{1}{\alpha+N}(\alpha H + \sum_{i=0}^N \delta_{\bar{\theta}_i}))$.

Stick-Breaking Construction. Infinite sequence of mixture weights $X = \{x_k\}_{k=1}^\infty$, $\beta_k \sim Beta(1,\alpha)$. $x_k = \beta_k(1 - \sum_{i=1}^{k-1} x_i)$. This is denoted as $X \sim GEM(\alpha)$.

1.22 Gaussian Processes

Problem setup. X_i input feature. $Y_i = f(X_i)$ output value. $f(\cdot)$ is an unknown function. Goal: given $D = \{(x_i, y_i)\}$ predict output y_* for new x_*.

Prediction. (1) Infer the posterior distribution of $f(x)$; (2) marginalize over $f(x)$ to obtain y_*.

$$\begin{aligned} &P(Y_*|X_{1:N}, Y_{1:N}, X_*) \\ &= \int P(f(x), Y_*|X_{1:N}, Y_{1:N}, X_*) df(x) \\ &= \int P(f(x)|X_{1:N}, Y_{1:N}) P(Y_*|f(x), X_*) df(x) \end{aligned}$$

Gaussian Process. Prior for $f(\cdot)$ w/o explicit parametrization. The distribution $P(\cdot)$ over $f(x)$ is a Gaussian Process if for any finite $\{x_1, \ldots, x_N\}$ the vector $f = [f(x_1), \ldots, f(x_N)]^T$ is Gaussian. $\mu = [E[f(x_1), \ldots] = [\mu(x_1), \ldots]$. $\mathbb{K} = E[(f-\mu)(f-\mu)^T]$ with $K_{ij} = K(x_i, x_j)$ PSD kernel Fn.

GP Regression. $f(\cdot) \sim GP(\mu(x), K(x,x'))$. (1) Noise free observations $Y_i = f(x_i) \triangleq f_i$, training data $D = \{(x_1, f_1), \ldots, (x_N, f_N)\}$. New observation x_*. $f = [f(x_1), \ldots]^T = [f_1, \ldots]^T \sim \mathcal{N}(\mu, \mathbb{K})$. $[f, f(x_*) = f_*]^T \sim \mathcal{N}([\mu, \mu_*]^T, [\mathbb{K}, k_*; k_*^T, k_*'])$ where $\mathbb{K} = K(x_{1:N}, x_{1:N})$, $k_* = K(x_{1:N}, x_*)$, $k_*' = K(x_*, x_*)$. Posterior $P(f_*|x_*, D) \sim \mathcal{N}(m, \sigma^2)$ where $m = E[f_*] + \Lambda_{*f}^T \Lambda_f^{-1}(f - \mu) = \mu_* + k_*^T \mathbb{K}^{-1}(f - \mu)$, $\sigma^2 = k_*' - k_*^T \mathbb{K}^{-1} k_*$. Generalize to multiple testing points $x_1^*, \ldots, x_M^*$, $[f, f_*] \sim \mathcal{N}([\mu, \mu_*]^T, [\mathbb{K}, \mathbb{K}_*; \mathbb{K}_*^T, \mathbb{K}_*'])$. $P(f|x_1^*, \ldots, x_M^*, D) = \mathcal{N}(f_*; m, \Sigma)$. $m = \mu_* + \mathbb{K}_*^T \mathbb{K}^{-1}(f - \mu)$. $\Sigma = K_*' - K_*^T K^{-1} K_*$.

(2) Noisy observations. $Y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \sim iid\ \mathcal{N}(0, \sigma_y^2)$. $E[Y_i] = \mu_i$, $Cov(Y_i, Y_j) = K(x_i, x_j) + \sigma_Y^2 \delta_{ij}$. So $Cov(Y_{1:N}) = K + \sigma_Y^2 I \triangleq K_Y$. Single observation $[Y_{1:N}, f_*]^T \sim \mathcal{N}([\mu, \mu_*]^T, [K_Y, k_*; k_*^T, k_*'])$. $p(f_*|x_*, D) = \mathcal{N}(f_*; m, \sigma^2)$ where $m = \mu_* + k_*^T K_Y^{-1}(y - \mu)$, $\sigma^2 = k_*' - k_*^T K_Y^{-1} k_*$.

Kernel Function. (prediction performance). RBF Kernel. $K(x,x') = \beta \exp\{-\frac{1}{2r^2}\|x - x'\|_F^2\}$. Or generalized form $K(x,x') = \beta \exp\{-\frac{1}{2r^2}(x-x')^T M(x-x')\}$ which can emphasize certain directions in data.

1.23 A. Useful Distributions

(1) Normal $X \sim \mathcal{N}(x;\mu,\sigma^2)$ (Continuous RV), $E[X] = \mu$, $Var[X] = \sigma^2$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\}$$

(2) Uniform $X \sim Unf(x;[a,b])$ (Continuous), $E[X] = \frac{a+b}{2}$, $Var[X] = \frac{(b-a)^2}{12}$.

$$f_X(x) = \frac{1}{b-a}\mathbb{I}\{x \in [a,b]\} + 0\mathbb{I}\{x \notin [a,b]\}$$

(3) Exponential $X \sim Exp(x;\lambda)$ (Continuous), $E[X] = 1/\lambda$, $Var[X] = 1/\lambda^2$.

$$f_X(x) = \lambda e^{-\lambda x} u(x)$$

(4) Bernoulli $X \sim B(x;p)$ (discrete RV), $E[X] = p$, $Var[X] = p(1-p)$.

$$f_X(x) = p^x(1-p)^{1-x}$$

(5) Poisson $X \sim Poisson(x;\mu)$ (discrete), $E[X] = \mu$, $Var[X] = \mu$.

$$f_X(x) = \frac{\mu^x}{x!}e^{-\mu}, x = 0,1,2,\ldots$$

(6) Beta $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ where $\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du$, namely $\Gamma(z+1) = z\Gamma(z)$. $E[X] = \frac{\alpha}{\alpha+\beta}$, $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

1.24 B. Tricks

$(AB)^{-1} = B^{-1}A^{-1}$.
A PSD means $x \in R^n \notin \{0\}$, $x^T Ax \geq 0$.
$E[Ax] = AE[x]$
$Var[Ax] = AVar[x]A^T$
$E[tr(AB)] = tr(E[AB])$
$\det|\alpha A| = \alpha^n A$
$tr(AB) = tr(BA)$
$|Cov(x,y)| \leq \sqrt{Var(x)Var(y)}$
Gaussian Integral $\int_{-\infty}^{+\infty} e^{-x^2}dx = \sqrt{\pi}$.
$\int_{-\infty}^{+\infty} ae^{-\frac{(x-b)^2}{2c^2}}dx = ac\sqrt{2\pi}$.
Integral by parts. $\int udv = uv - \int vdu$.
steady markov: $\pi = A\pi$.