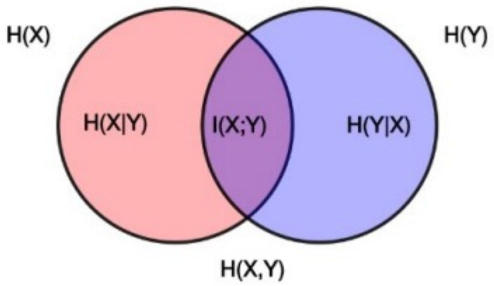
Equation Sheet (W.I.P)

Maximum Likelihood Estimation and Logistic Regression

- **Odds** : $odds(p) = \frac{p}{1-p}$
- **Logit (Logarithms of the Odds)** : $Logit(p) = \log(\frac{p}{1-p}) = -\log(\frac{1}{p} - 1)$
- **Linear Regression** : $\hat{y} = \theta_0 x_0 + \sum_{k=1}^{K} \theta_k x_{ik} + \epsilon_i = \sum_{k=0}^{K} \theta_k x_{ik} + \epsilon_i = \theta^T x_i + \beta_i$
- **Logistic Regression**:
  - $h_\theta(X) = P(Y=1|X;\theta) = \frac{1}{1+exp(-\theta^T X)} = \frac{1}{1+\frac{1}{odds(p)}}$
  - $P(Y=0|X;\theta) = 1 - h_\theta(X)$
  - Bernoulli Distribution : $P(y|X;\theta) = \text{Bernoulli}(h_\theta(X)) = h_\theta(X)^y \times (1-h_\theta(X))^{1-y}$
  - Likelihood Function of Bernoulli Distribution: $L(\theta|y;x) = P(Y|X;\theta) = \prod_i P(y_i|x_i;\theta) = \prod_i h_\theta(x_i)^{y_i} \times (1-h_\theta(x_i))^{1-y_i}$
  - Cost Function of Likelihood Function : $-\log(L(\theta|y;x)) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(h_\theta(x_i)) + (1-y_i)\log(1-h_\theta(x_i)))$
  - Maximum Likelihood Function : $\hat{\theta}_{\text{MLE}} = \text{argmin}_\theta(-\log(L(\theta|y;x)))$

Information Theory

- **Self-Information** : $I_x(x) = -log_b[P_X(x)] = \log_b(\frac{1}{P_X(x)})$
  - b is the unit we want the information to be in
  - Relates to logit : $logit(x) = I(\neg x) - I(x)$
- **Entropy** : $H(X) \equiv E[I_X(x)] \equiv -\sum_i^n P(X=x_i) \times \log_b(P(X=x_i)) \equiv E[\log_b \frac{1}{P_X(x)}] \equiv -E[\log_b P_X(x)]$
  - n is the number of independent variables



- **Joint Entropy** : $H(X,Y) = -E[\log p(X,Y)] = -\sum_{x_i \in R_X}\sum_{y_j \in R_Y} p(x_i,y_j)\log p(x_i,y_j)$
- **Conditional Entropy** : $H(Y,X) = -E[\log p(Y|X)] = -\sum_{x_i \in R_X}\sum_{y_j \in R_Y} p(x_i,y_j)\log p(y_j|x_i) = H(X,Y) - H(X)$
  - Conditional Probability : $p(x|y) = \frac{p(x,y)}{p(x)}$
  - $H(Y|X) = H(X,Y) - H(X)$
  - $H(X|Y) = H(X,Y) - H(Y)$
- **Relative Entropy (Kullback-Leibler Divergence)** : $D_{KL}(P||Q) = \sum_{x \in R_X} P(x)\log\frac{P(x)}{Q(x)}$
  - **Cross Entropy** - $H(P||Q) = -\sum_{x \in R_X} P(x)\log Q(x) = H(P) + D_{KL}(P||Q)$
  - **Jensen-Shannon Divergence** (JSD)- $JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$
- **Mutual Information** : $I(X;Y) = \sum_{x \in R_X}\sum_{y \in R_Y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$
  - Can be defined in KL Divergence: $I(X;Y) = D_{KL}(P(X,Y)||P(X)P(Y))$
    - X and Y are independent if and only if $I(X;Y) = 0$

Decision Trees

- **Gini Index/Impurity** : $I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$
  - p_i is the fraction of the number of items of class i over the total number of items
- **Information Gain** : $IG(Y,X) = H(Y) - H(Y|X) = I(X;Y)$

Bayes' Theorem and Bayesian Network 1

- **Bayes' Theorem** : $P(A|B) = \frac{P(B|A)\times P(A)}{P(B)}$
- **Likelihood Function of the observed variable given different values of** Θ : $
$P(X=n|\Theta) = p(x|\Theta) = \{p(x_n|\theta_1), p(x_n|\theta_2), \ldots, p(x_n|\theta_m)\}$
- **Bayes' Theorem for Discrete Distribution** : $p(\Theta|x) = \frac{p(x|\Theta)\times p(\Theta)}{p(x)}$
  - $p(\Theta|x)$ is the **posterior** (based on known knowledge) distribution
  - $p(x|\Theta)$ is the **Likelihood function** (where x is the value of X and $x \in R_X$)
  - $p(x)$ is the **marginal likelihood**
- **Full Joint Distribution** : $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))$
- **Conditional Independence** - $(A \perp\!\!\!\perp B)C \iff P(A, B|C) = P(A|C) \times P(B|C)$
- **Relationships Joint Distribution**
  - Direct Cause - $P(W|R)$
  - Indirect Cause - $P(C, R, W) = P(C) \times P(R|C) \times P(W|R)$
  - Common Effect - $P(S, R, W) = P(S) \times P(R) \times P(W|S, R)$
  - Common Cause $P(C, S, R) = P(C) \times P(S|C) \times P(R|C)$