Possibly Useful Formulas

Probability:

$$P(X=x)=\Pr(X=x) \text{ or } P(X=x)=\frac{d}{dx}\Pr(X\le x)$$

$$P(X,Y)=P(X|Y)P(Y)$$

$$E\left[f(X,Y)\right]=\sum_{x,y} f(x,y)P(X=x,Y=y)$$

MAP Decision: $f(x)=\underset{y}{\operatorname{argmax}} P(Y=y|X=x)$

Bayes Error Rate: $=\sum_x P(X=x)\min_y P(Y\neq y|X=x)$

Precision $=P(Y=1|f(X)=1)=\dfrac{TP}{TP+FP}$

Recall=Sensitivity $=P(f(X)=1|Y=1)=\dfrac{TP}{TP+FN}$

Specificity $=P(f(X)=0|Y=0)=\dfrac{TN}{TN+FP}$

Naive Bayes: $f(x)\approx \underset{y}{\operatorname{argmax}}\left(\ln P(Y=y)+\sum_{i=1}^{n}\ln P(W=w_i|Y=y)\right)$

Laplace Smoothing: $P(W=w_i|Y=y)=\dfrac{k+\text{Count}(w_i,y)}{k+\sum_{v\in\mathscr{V}}\left(k+\text{Count}(v,y)\right)}$

HMM: $v_1(j)=\pi(j)b_j(\mathbf{x}_t)$

$$v_t(j)=\max_i v_{t-1}(i)a_{i,j}b_j(\mathbf{x}_t),\ \ \psi_t(j)=\underset{i}{\operatorname{argmax}}\, v_{t-1}(i)a_{i,j}b_j(\mathbf{x}_t)$$

$$y^*(T)=\underset{i}{\operatorname{argmax}}\, v_T(i),\ \ y^*(t)=\psi_{t+1}(y^*(t+1))$$

Demographic Parity: $P(f(X)|A=1)=P(f(X)|A=0)$

Equal Odds: $P(f(X)|Y,A=1)=P(f(X)|Y,A=0)$

Predictive Parity: $P(Y|f(X),A=1)=P(Y|f(X),A=0)$

Learning: $\mathscr{R}=E\left[\ell(Y,f(X))\right],\ \ \mathscr{R}_{emp}=\dfrac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i))$

Linear Regression : $f(\mathbf{x})=\mathbf{w}^T\mathbf{x}+b$

$$\mathscr{L}=\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}_i,\ \ \mathscr{L}_i=\frac{1}{2}\varepsilon_i^2,\varepsilon_i=f(\mathbf{x}_i)-y_i$$

Linear Classifier: $f(\mathbf{x}) = \text{argmax}\,(\mathbf{W}\mathbf{x}+\mathbf{b})$

Perceptron: $\mathbf{w}_c \leftarrow \begin{cases} \mathbf{w}_c - \eta\mathbf{x} & c = \text{argmax}\,(\mathbf{W}\mathbf{x}+\mathbf{b}) \\ \mathbf{w}_c + \eta\mathbf{x} & c = y \\ \mathbf{w}_c & \text{otherwise} \end{cases}$

Softmax: $f_c(\mathbf{x}) = \dfrac{\exp(\mathbf{w}_c^T\mathbf{x}+b_c)}{\sum_{k=1}^{|\mathcal{Y}|}\exp(\mathbf{w}_k^T\mathbf{x}+b_k)} \approx P(Y=c|\mathbf{x})$

Sigmoid: $\sigma(\mathbf{w}^T\mathbf{x}+b) = \dfrac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}} \approx P(Y=1|\mathbf{x})$

Cross Entropy: $\mathscr{L} = -\ln f_y(\mathbf{x}),\ \dfrac{\partial\mathscr{L}}{\partial f_c(\mathbf{x})} = \begin{cases} -\frac{1}{f_c(\mathbf{x})} & c = y \\ 0 & \text{otherwise} \end{cases}$

SGD: $\mathbf{w}_C \leftarrow \mathbf{w}_c - \eta\dfrac{\partial\mathscr{L}}{\partial\mathbf{w}_c} = \begin{cases} \mathbf{w}_c - \eta(f_c(\mathbf{x}_i)-1)\mathbf{x}_i & c = y \\ \mathbf{w}_c - \eta(f_c(\mathbf{x}_i)-0)\mathbf{x}_i & \text{otherwise} \end{cases}$

Pinhole Camera: $\dfrac{x'}{f} = -\dfrac{x}{z},\ \dfrac{y'}{f} = -\dfrac{y}{z}$

Image Gradient:

$$h_x(x',y') = \frac{h(x'+1,y')-h(x'-1,y')}{2},\ h_y(x',y') = \frac{h(x',y'+1)-h(x',y'-1)}{2}$$

Convolution:

$$y[k,l] = \sum_i\sum_j x[k-i,l-j]h[i,j],\ \frac{\partial y[k,l]}{\partial h[i,j]} = x[k-i,l-j]$$

Max Pooling:

$$z[m] = \max_{(m-1)p+1\le k\le mp} y[k],\ \frac{\partial z[m]}{\partial y[j]} = \begin{cases} 1 & j = \text{argmax}_{(m-1)p+1\le k\le mp}\, y[k] \\ 0 & \text{otherwise} \end{cases}$$

Admissible: $\hat{h}(n) \le h(n)$

Consistent: $\hat{h}(n) - \hat{h}(m) \le h(n,m)$

Value Iteration: $u_i(s) = r(s) + \gamma \max_a \sum_{s'} P(s'|s,a) u_{i-1}(s')$

Policy Evaluation: $u_\pi(s) = r(s) + \gamma \sum_{s'} P(s'|s,\pi(s)) u_\pi(s')$

Policy Improvement: $\pi_{i+1}(s) = \operatorname{argmax}_a \sum_{s'} P(s'|s,a) u_{\pi_i}(s')$

Alpha-Beta Max Node: $v = \max(v, \text{child}); \quad \alpha = \max(\alpha, \text{child})$

Alpha-Beta Min Node: $v = \min(v, \text{child}); \quad \beta = \min(\beta, \text{child})$

Expectiminimax: $u(s) = \begin{cases} \max_a \sum_{s'} P(s'|a,a) u(s') & s \in \text{max states} \\ \min_a \sum_{s'} P(s'|a,a) u(s') & s \in \text{min states} \end{cases}$

Mixed Nash Equilibrium: $\begin{array}{l} P(A=0) r_B(0,0) + P(A=1) r_B(1,0) = P(A=0) r_B(0,1) + P(A=1) r_B(1,1) \\ P(B=0) r_A(0,0) + P(B=1) r_A(0,1) = P(B=0) r_A(1,0) + P(B=1) r_A(1,1) \end{array}$

Unification: $S : \{\mathscr{V}_P, \mathscr{V}_Q\} \to \{\mathscr{V}_Q, \mathscr{C}\}$ such that $S(P) = S(Q) = U$

CBOW Generative: $\mathscr{L} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{j=-c, j\neq 0}^{c} \ln \frac{\exp(\mathbf{v}_t^T \mathbf{v}_{t_j})}{\sum_{\mathbf{v} \in \mathscr{V}} \exp(\mathbf{v}^T \mathbf{v}_{t+j})}$

Skip-gram Contrastive: $\mathscr{L} = -\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{\mathbf{v}' \in \mathscr{D}_+(w_t)} \ln \frac{1}{1 + e^{-\mathbf{v}'^T \mathbf{v}_t}} + \sum_{\mathbf{v}' \in \mathscr{D}_-(w_t)} \ln \frac{1}{1 + e^{\mathbf{v}'^T \mathbf{v}_t}} \right)$

Transformer: $\mathbf{c}_t = \sum_s \alpha(t,s) \mathbf{v}_s$

Attention: $\alpha(t,s) = \frac{\exp(\mathbf{q}_t^T \mathbf{k}_s)}{\sum_{s'} \exp(\mathbf{q}_t^T \mathbf{k}_{s'})}$

Model-based Learning: $P(s_{t+1}|s_t, a_t) = \frac{\text{Count}(s_t, a_t, s_{t+1}) + k}{\sum_{s' \in \mathscr{S}} (\text{Count}(s_t, a_t, s') + k)}$

On-policy learning: $\mathbf{W}_{a_t} \leftarrow \mathbf{W}_{a_t} + \eta \nabla_{\mathbf{W}_{a_t}} \ln P(\mathbf{s}_{t+1}|\mathbf{s}_t, a_t)$

Off-policy learning: $\mathbf{W} \leftarrow \mathbf{W} + \eta \nabla_{\mathbf{W}} \ln P(\mathbf{s}_{t+1}|\mathbf{s}_t, a_t)$

Epsilon-first: $N_{first} = \frac{1}{\varepsilon}$

Epsilon-greedy: If $z \leq \varepsilon$ then explore, $z \in (0,1)$

TD-learning: $q_{local}(s_t, a_t) = r_t + \gamma \max_{a' \in \mathscr{A}} q_t(s_{t+1}, a')$

SARSA: $q_{local}(s_t, a_t) = r_t + \gamma q_t(s_{t+1}, a_{t+1})$

Q-learning: $q_{t+1}(s_t, a_t) = q_t(s_t, a_t) + \eta (q_{local}(s_t, a_t) - q_t(s_t, a_t))$

Deep Q: $\theta_{t+1} = \theta_t - \eta \frac{\partial}{\partial \theta} \frac{1}{2} (q_t(s_t, a_t) - q_{local}(s_t, a_t))^2$

Policy Gradient: $\frac{\partial u(s_t)}{\partial \theta} = \sum_\tau \frac{\partial P(\tau)}{\partial \theta} v(\tau) = \sum_\tau P(\tau) \frac{\partial \ln P(\tau)}{\partial \theta} v(\tau) = E \left[\frac{\partial \ln P(\tau)}{\partial \theta} v(\tau) \right]$

Standard error: $M = \frac{1}{n} \sum_{i=1}^{n} Y_i, \ \text{stdev}(M) = \frac{\sigma}{\sqrt{n}}$