KNN Find k examples $\{\mathbf{x}^{(i)}, t^{(i)}\}$ closest to the test instance $\mathbf{x}$ and then output majority $\arg\max_{t^z} \sum_{r=1}^{k} \delta(t^{(z)}, t^{(r)})$. Define $\delta(a,b)=1$ if $a=b$, 0 otw. **Choice of k:** Rule is $k<\sqrt{n}$, small k may overfit, while large may underfit. **Curse of Dim:** In high dimensions, "most" points are approximately the same distance. **Computation Cost:** 0 (minimal) at trianing/ no learning involved. Query time find N distances in D dimension $\mathcal{O}(ND)$ and $\mathcal{O}(N\log N)$ sorting time.

Entropy $H(X) = -\mathbb{E}_{X\sim p}\left[\log_2 p(X)\right] = -\sum_{x\in X} p(x)\log_2 p(x)$ **Multi-class:** $H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y)$ **Properties:** H is non-negative, $H(Y|X)\leq H(Y)$, $X\perp Y \implies H(Y|X)=H(Y)$, $H(Y|Y)=0$, and $H(X,Y)=H(X|Y)+H(Y)=H(Y|X)+H(X)$

Expected Conditional Entropy $H(Y|X) = \mathbb{E}_{X\sim p(x)}[H(Y|X)] = \sum_{x\in X} p(x)H(Y|X=x) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x) = -\mathbb{E}_{(X,Y)\sim p(x,y)}\left[\log_2 p(Y|X)\right]$

Information Gain $IG(Y|X) = H(Y) - H(Y|X)$

Bias Variance Decomposition Using the square error loss $L(y,t)=\frac{1}{2}(y-t)^2$, **Bias ($\uparrow \implies$ underfitting):** How close is our classifier to true target. **Variance ($\uparrow \implies$ overfitting):** How widely dispersed are out predictions as we generate new datasets

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(h_{\mathcal{D}}(\mathbf{x})-t\right)^2\right] &= \mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]+\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]-t\right)^2\right] \\
&= \mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]\right)^2+\left(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]-t\right)^2+2\left(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]\right)\left(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]-t\right)\right] \\
&= \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]\right)^2\right]}_{\text{variance}}+\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(\mathbf{x})\right]-t\right)^2\right]}_{\text{bias}}
\end{aligned}$$