CSC311: Introduction to Machine Learning

Tianyu Du

October 19, 2019

Contents

1 Decision Trees

1.1 Entropy: An Information Theoretical Metric

Definition 1.1. Accuracy gain from splitting R into R_1 and R_2 based on loss $L(R)$:

$$L(R) - \underbrace{\frac{|R_1|L(R_1) + |R_2|L(R_2)}{|R_1| + |R_2|}}_{\text{New Loss}} \tag{1.1}$$

Typically, $L(R) = 1 - \mathbb{1}\{\% \text{ correct classified}\}$.

Definition 1.2. Given a random variable $X \sim p$, the **entropy** measures the amount of randomness/uncertainty in an arbitrary realization of X.

$$H(X) := \mathbb{E}_{X\sim p}[-\log_2 p(X)] \tag{1.2}$$

Proposition 1.1. Uniform random variable has the highest entropy.

Remark 1.1. Entropy is defined using $\log_2$ instead of ln because of an information theory perspective. In actual implementation of optimization algorithms, ln is used to derive gradient ascent rules. These two definitions are equivalent as optimization is invariant under positive monotone transformation.

Definition 1.3. Given joint distribution $(X,Y) \sim p(X,Y)$, the **entropy of joint distribution** is defined as

$$\begin{aligned} H(X,Y) &:= \mathbb{E}_{(X,Y)\sim p(X,Y)}[-\log_2 p(X,Y)] && (1.3)\\ &= -\sum_{x\in Im(X)} \sum_{y\in Im(Y)} p(x,y)\log_2 p(x,y) && (1.4) \end{aligned}$$

Definition 1.4. Given two random variables X and Y, the **conditional entropy of Y conditioned on specific realization of X** is defined to be

$$\begin{aligned} H(Y|X=x) &:= \mathbb{E}_{y\sim p(y|X=x)}[-\log_2 p(y|X=x)] && (1.5)\\ &= -\sum_{y\in Im(Y)} p(y|X=x)\log_2 p(y|X=x) && (1.6) \end{aligned}$$

The **expected conditional entropy**[1] is defined as

$$\begin{aligned} H(Y|X) &= \mathbb{E}_{X\sim p(x)}[H(Y|X)] && (1.7)\\ &= \mathbb{E}_{X\sim p(x)}[\mathbb{E}_{y\sim p(y|X=x)}[-\log_2 p(y|X=x)]] && (1.8)\\ &= \sum_{x\in X} p(x)H(Y|X=x) && (1.9)\\ &= -\sum_{x\in X} p(x) \sum_{y\in Im(Y)} p(y|X=x)\log_2 p(y|X=x) && (1.10)\\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|X=x) && (1.11)\\ &= -\mathbb{E}_{(X,Y)\sim p(x,y)}[\log_2 p(Y|X)] && (1.12) \end{aligned}$$

[1] This is independent of specific realization of X

Proposition 1.2. For every $X \in \Delta(\mathcal{X})$, $H(X) \geq 0$.

Proof. Immediate. ∎

Proposition 1.3 (Chain Rule)**.**

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) \tag{1.13}$$

Proof.

$$\begin{align}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \tag{1.14}\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 (p(x|y)p(y)) \tag{1.15}\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)[\log_2 p(x|y) + \log_2 p(y)] \tag{1.16}\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x|y) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y) \tag{1.17}\\
&= H(X|Y) - \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log_2 p(y) \tag{1.18}\\
&= H(X|Y) - \sum_{y\in\mathcal{Y}}\left\{\log_2 p(y)\sum_{x\in\mathcal{X}} p(x,y)\right\} \tag{1.19}\\
&= H(X|Y) - \sum_{y\in\mathcal{Y}}(\log_2 p(y))p(y) \tag{1.20}\\
&= H(X|Y) + H(Y) \tag{1.21}
\end{align}$$

∎

Proposition 1.4. If $X \perp Y$, then knowing X does not provide extra information (i.e. reduce entropy) of Y. That is $H(Y|X) = H(Y)$.

Proof.

$$\begin{align}
H(Y|X) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|X=x) \tag{1.22}\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y) \tag{1.23}\\
&= -\sum_{y\in\mathcal{Y}}\log_2 p(y)\sum_{X\in\mathcal{X}} p(x,y) \tag{1.24}\\
&= -\sum_{y\in\mathcal{Y}} p(y)\log_2 p(y) \tag{1.25}\\
&= H(Y) \tag{1.26}
\end{align}$$

∎

Proposition 1.5. Y becomes deterministic by knowing Y, that is, $H(Y|Y) = 0$.

Proof.

$$\begin{aligned} H(Y|Y) &= -\sum_{y\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} p(y,y)\log_2 p(y|Y=y) \tag{1.27}\\ &= -\sum_{y\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} p(y,y)\log_2 1 \tag{1.28}\\ &= 0 \tag{1.29}\end{aligned}$$

■

Proposition 1.6. By knowing X, the uncertainty about Y is reduced: $H(Y|X) \leq H(Y)$.

Proof. TODO: *Proof this* ■

Definition 1.5. The **information gain** in Y due to X, or **mutual information** of X and Y is defined to be

$$IG(Y|X) := H(Y) - H(Y|X) \tag{1.30}$$

When X is completely uninformative about Y: $H(Y|X) = H(Y)$, then $IG(Y|X) = 0$.
When X is completely information about Y: $H(Y|X) = 0$ (deterministic), then $IG(Y|X) = H(Y)$.

Proposition 1.7 (Symmetry of Information Gain)**.**

$$\begin{aligned} IG(Y|X) &:= H(Y) - H(Y|X) \tag{1.31}\\ &= H(X,Y) - H(X|Y) - H(Y|X) \tag{1.32}\\ &= H(Y|X) + H(X) - H(X|Y) - H(Y|X) \tag{1.33}\\ &= H(X) - H(X|Y) \tag{1.34}\\ &= IG(X|Y) \tag{1.35}\end{aligned}$$

2 Decision Trees

Remark 2.1. Greedy algorithms don't necessarily yield the global optimum.

Definition 2.1. An **ensemble** of predictors is a set of predictors whose individual decisions are combined in some way to predict new examples

3 Bias-Variance Decomposition

Let $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathbb{R}$ denote one training instance such that

$$(\mathbf{x}^{(i)}, y^{(i)}) \overset{i.i.d.}{\sim} p_{\text{sample}} \tag{3.1}$$

where $p_{\text{sample}} \in \Delta(\mathcal{X} \times \mathbb{R})$.
Fixing $N \in \mathbb{N}$, one can construct a new distribution $p_{\text{dataset}} \in \Delta(\mathcal{X} \times \mathbb{R})^N$ such that

$$(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^{N} =: \mathcal{D} \sim p_{\text{dataset}} \tag{3.2}$$

Given a (random) training set $\mathcal{D}$, a (random) classifier function $h_{\mathcal{D}} \in \mathcal{H}$ is generated.
For every *query point* $\mathbf{x} \in \mathcal{X}$, the prediction $h_{\mathcal{D}}(\mathbf{x})$ is therefore random.

Suppose y is not deterministic in x, then the expected mean squared error when the model is applied on new instances sampled from p_{sample} is

$$\begin{align}
\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x})-y)^2] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x},y}[(h_{\mathcal{D}}(\mathbf{x})-y)^2|\mathcal{D}]] \tag{3.3}\\
&= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x},y}[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x]+\mathbb{E}_y[y|x]-y)^2|\mathcal{D}]] \tag{3.4}\\
&= \mathbb{E}_{\mathcal{D}}\{\mathbb{E}_x[\mathbb{E}_y[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])^2]] \tag{3.5}\\
&\quad + 2\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])(\mathbb{E}_y[y|x]-y)] \tag{3.6}\\
&\quad + \mathbb{E}_{x,y}(\mathbb{E}_y[y|x]-y)^2\} \tag{3.7}\\
&= \mathbb{E}_{\mathcal{D}}\{\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])^2] \tag{3.8}\\
&\quad + 2\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])(\mathbb{E}_y[y|x]-y)] \tag{3.9}\\
&\quad + \mathbb{E}_{x,y}[(\mathbb{E}_y[y|x]-y)^2]\} \tag{3.10}\\
& \tag{3.11}
\end{align}$$

By law of iterative expectation,

$$\begin{align}
\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])(\mathbb{E}_y[y|x]-y)] &= \mathbb{E}_x[\mathbb{E}_y[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])(\mathbb{E}_y[y|x]-y)]] \tag{3.12}\\
&= \mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])(\mathbb{E}_y[y|x]-\mathbb{E}_y[y])] \tag{3.13}\\
&= 0 \tag{3.14}
\end{align}$$

By dropping irrelevant expectation operators,

$$\begin{align}
\Delta &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])^2]] + \underbrace{\mathbb{E}_{x,y}[(\mathbb{E}_y[y|x]-y)^2]}_{\text{Bayes Error } \varepsilon^2} \tag{3.15}\\
&= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_y[y|x])^2]] + \varepsilon^2 \tag{3.16}\\
&= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x]+\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x]-\mathbb{E}_y[y|x])^2]] + \varepsilon^2 \tag{3.17}
\end{align}$$

Note that

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x])(\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x]-\mathbb{E}_y[y|x])]] = 0 \tag{3.18}$$

The first component reduced to zero after applying law of iterative expectation.
Therefore,

$$\begin{align}
\Delta &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}(h_{\mathcal{D}}(x)|x))^2]] + \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x]-\mathbb{E}_y[y|x])^2]] + \varepsilon^2 \tag{3.19}\\
&= \underbrace{\mathbb{E}_{x,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x})-\mathbb{E}_{\mathcal{D}}(h_{\mathcal{D}}(x)|x))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_x[(\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x]-\mathbb{E}_y[y|x])^2]}_{\text{Bias Squared}} + \varepsilon^2 \tag{3.20}\\
& \tag{3.21}
\end{align}$$

4 Bagging

5 Bayes Optimality

Theorem 5.1.

$$\operatorname*{argmin}_{y} \mathbb{E}[(y-t)^2|\mathbf{x}] = \mathbb{E}[t|\mathbf{x}] \tag{5.1}$$

where $t \sim p(t|\mathbf{x})$.