1 Theory

1.1 Information Entropy

Definition 1.1. **Accuracy gain** from splitting R into R_1 and R_2 based on loss $L(R)$: $L(R) - \frac{|R_1|L(R_1)+|R_2|L(R_2)}{|R_1|+|R_2|}$

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

$$H(X) := \mathbb{E}_{X\sim p}[-\log_2 p(X)] \tag{1.1}$$

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

$$H(X, Y) := \mathbb{E}_{(X,Y)\sim p(X,Y)}[-\log_2 p(X, Y)] = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y) \tag{1.2}$$

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

$$H(Y|X = x) := \mathbb{E}_{y\sim p(y|X=x)}[-\log_2 p(y|X = x)] = -\sum_{y\in\mathcal{Y}} p(y|X = x)\log_2 p(y|X = x) \tag{1.3}$$

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

$$H(Y|X) = \mathbb{E}_{X\sim p(x)}[H(Y|X)] = \mathbb{E}_{X\sim p(x)}[\mathbb{E}_{y\sim p(y|X=x)}[-\log_2 p(y|X = x)]] = \sum_{x\in\mathcal{X}} p(x)H(Y|X = x) \tag{1.4}$$

$$= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in Im(Y)} p(y|X = x)\log_2 p(y|X = x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|X = x) = -\mathbb{E}_{(X,Y)\sim p(x,y)}[\log_2 p(Y|X)] \tag{1.5}$$

Proposition 1.1. For every $X \in \Delta(\mathcal{X})$, $H(X) \geq 0$.

Proposition 1.2 (Chain Rule)**.** $H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$

Proposition 1.3. If $X \perp Y$, then knowing X does not provide extra information (i.e. reduce entropy) of Y. That is $H(Y|X) = H(Y)$.

Proposition 1.4. Y becomes deterministic by knowing Y, that is, $H(Y|Y) = 0$.

Proposition 1.5. By knowing X, the uncertainty about Y is reduced: $H(Y|X) \leq H(Y)$.

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

$$IG(Y|X) := H(Y) - H(Y|X) \tag{1.6}$$

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 1.6 (Symmetry of Information Gain)**.**

$$IG(Y|X) := H(Y) - H(Y|X) = H(X, Y) - H(X|Y) - H(Y|X) \tag{1.7}$$

$$= H(Y|X) + H(X) - H(X|Y) - H(Y|X) = H(X) - H(X|Y) = IG(X|Y) \tag{1.8}$$

BVD: Deterministic

$$\mathbb{E}_{x,\mathcal{D}}\left[(h_{\mathcal{D}}(x) - f(x))^2\right] = \mathbb{E}_{x,\mathcal{D}}\left[(h_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(x)|x\right])^2\right] + \mathbb{E}_x\left[(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(x)|x\right] - f(x))^2\right] \tag{1.9}$$

BVD: Stochastic Let $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X}\times\mathbb{R}$ denote one training instance such that $(\mathbf{x}^{(i)}, y^{(i)}) \overset{i.i.d.}{\sim} p_{\text{sample}}$,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 $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^N =: \mathcal{D} \sim p_{\text{dataset}}$ 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

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x},y}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2|\mathcal{D}]] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x},y}[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x] + \mathbb{E}_y[y|x] - y)^2|\mathcal{D}]] \tag{1.10}$$

$$= \mathbb{E}_{\mathcal{D}}\{\mathbb{E}_x[\mathbb{E}_y[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])^2]] + 2\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)] + \mathbb{E}_{x,y}(\mathbb{E}_y[y|x] - y)^2\} \tag{1.11}$$

$$= \mathbb{E}_{\mathcal{D}}\{\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])^2] + 2\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)] + \mathbb{E}_{x,y}[(\mathbb{E}_y[y|x] - y)^2]\} \tag{1.12}$$

$$\tag{1.13}$$

By law of iterative expectation,

$$\mathbb{E}_{x,y}[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)] = \mathbb{E}_x[\mathbb{E}_y[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - y)]] = \mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])(\mathbb{E}_y[y|x] - \mathbb{E}_y[y])] = 0 \tag{1.14}$$

[1] This is independent of specific realization of X

By dropping irrelevant expectation operators,

$$\Delta = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])^2]] + \underbrace{\mathbb{E}_{x,y}[(\mathbb{E}_y[y|x] - y)^2]}_{\text{Bayes Error } \varepsilon^2} = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_y[y|x])^2]] + \varepsilon^2 \tag{1.15}$$

$$= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x] + \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x] - \mathbb{E}_y[y|x])^2]] + \varepsilon^2 \tag{1.16}$$

Note that $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_x[(h_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x])(\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)|x] - \mathbb{E}_y[y|x])]] = 0$ The first component reduced to zero after applying law of iterative expectation. Non-deterministic case

$$\mathbb{E}_{x,y,\mathcal{D}}\left[(h_{\mathcal{D}}(x) - y)^2\right] = \mathbb{E}_{x,\mathcal{D}}\left[(h_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(x)|x\right])^2\right] + \mathbb{E}_x\left[(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(x)|x\right] - \mathbb{E}_y[y|x])^2\right] + \mathbb{E}_{x,y}\left[(\mathbb{E}_y[y|x] - y)^2\right] \tag{1.17}$$

2 Mathematics & Probability

$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$
$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right] \in \mathbb{R}^{d\times d}$
$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_y)^T\right] \in \mathbb{R}^{d\times d}$ $\quad p(\theta|\text{ data }) = \frac{p(\text{ data }|\theta)p(\theta)}{p(\text{ data })}$ $\quad \theta^{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta|\text{ data }) = \underset{\theta}{\operatorname{argmax}} p(\text{data }|\theta)p(\theta)$

$\theta^{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}} p(X_1,\ldots,X_N|\theta)\,p(\theta) = \underset{\theta}{\operatorname{argmax}} p(\theta)\prod_{i=1}^{N} p(X_i|\theta) = \underset{\theta}{\operatorname{argmax}} \log p(\theta) + \sum_{i=1}^{N}\log p(X_i|\theta)$

Proposition 2.1 (Law of Total Expectation)**.** $\mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] = \mathbb{E}[X]$.

Proof. $\mathbb{E}[\mathbb{E}[X|Y]] = \int\left[\int xp(x|y)dx\right]p(y)dy = \iint xp(x,y)dxdy = \mathbb{E}[X]$ ■

$\underset{\mathbf{w}}{\operatorname{minimize}}\mathcal{J}(\mathbf{w}) =: \frac{1}{2}\|\mathbf{t}-\mathbf{X}\mathbf{w}\|_2^2 \quad \mathcal{J}(\mathbf{w}) = \frac{1}{2}\|\mathbf{t}\|_2^2 + \frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{t}^\top\mathbf{X}\mathbf{w}$.

Theorem 2.1 (Bayes Optimal)**.** $\operatorname{argmin}_y \mathbb{E}[(y-t)^2|\mathbf{x}] = \mathbb{E}[t|\mathbf{x}]$ where $t \sim p(t|\mathbf{x})$.

Multi-class Classification $\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}$ (aka *logits*) Input dim $= D$, output dim $= K$, $\mathbf{W} \in \mathbb{R}^{K\times D}$
Pred_prob:$y_k = \operatorname{softmax}(z_1,\ldots,z_K)_k = \frac{e^{z_k}}{\sum_{k'}e^{z_{k'}}}$ $\quad \mathcal{L}_{\mathrm{CE}}(\mathbf{y},\mathbf{t}) = -\sum_{k=1}^{K} t_k\log y_k = -\mathbf{t}^T(\log(y))$ (*Softmax-cross-entropy*).

$$\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\mathbf{w}_k} = \frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial z_k}\cdot\frac{\partial z_k}{\mathbf{w}_k} = (y_k - t_k)\cdot\mathbf{x}, \quad \mathbf{w}_k \leftarrow \mathbf{w}_k - \alpha\frac{1}{N}\sum_{i=1}^{N}\left(y_k^{(i)} - t_k^{(i)}\right)\mathbf{x}^{(i)}, \quad \mathbf{W} \leftarrow \mathbf{W} - \frac{\alpha}{N}(\mathbf{y}-\mathbf{t})\mathbf{X}$$

(Forward)

$$\mathbf{z} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \quad \mathbf{h} = \sigma(\mathbf{z}) \quad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \quad \mathcal{L} = \tfrac{1}{2}\|\mathbf{t}-\mathbf{y}\|^2$$

(Backward)

$$\begin{aligned}\overline{\mathcal{L}} &= 1\\ \overline{\mathbf{y}} &= \overline{\mathcal{L}}(\mathbf{y}-\mathbf{t})\\ \overline{\mathbf{W}^{(2)}} &= \overline{\mathbf{y}}\mathbf{h}^\top\\ \overline{\mathbf{h}^{(2)}} &= \overline{\mathbf{y}}\\ \overline{\mathbf{h}} &= \mathbf{W}^{(2)\top}\overline{\mathbf{y}}\\ \overline{\mathbf{z}} &= \overline{\mathbf{h}}\circ\sigma'(\mathbf{z})\\ \overline{\mathbf{W}^{(1)}} &= \overline{\mathbf{z}}\mathbf{x}^\top\\ \overline{\mathbf{b}^{(1)}} &= \overline{\mathbf{z}}\end{aligned}$$

(Forward: Reg)

$$\begin{aligned}z &= wx + b\\ y &= \sigma(z)\\ \mathcal{L} &= \frac{1}{2}(y-t)^2\\ \mathcal{R} &= \frac{1}{2}w^2\\ \mathcal{L}_{\mathrm{reg}} &= \mathcal{L} + \lambda\mathcal{R}\end{aligned}$$

(Backward I)

$$\begin{aligned}\overline{\mathcal{L}_{\mathrm{reg}}} &= 1\\ \overline{\mathcal{R}} &= \overline{\mathcal{L}_{\mathrm{reg}}}\frac{\mathrm{d}\mathcal{L}_{\mathrm{reg}}}{\mathrm{d}\mathcal{R}}\\ &= \overline{\mathcal{L}_{\mathrm{reg}}}\lambda\\ \overline{\mathcal{L}} &= \overline{\mathcal{L}_{\mathrm{reg}}}\frac{\mathrm{d}\mathcal{L}_{\mathrm{reg}}}{\mathrm{d}\mathcal{L}}\\ &= \overline{\mathcal{L}_{\mathrm{reg}}}\\ \bar{y} &= \overline{\mathcal{L}}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}y}\\ &= \overline{\mathcal{L}}(y-t)\end{aligned}$$

(Backward II)

$$\begin{aligned}\bar{z} &= \bar{y}\frac{\mathrm{d}y}{\mathrm{d}z}\\ &= \bar{y}\sigma'(z)\\ \bar{w} &= \bar{z}\frac{\partial z}{\partial w} + \overline{\mathcal{R}}\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}w}\\ &= \bar{z}x + \overline{\mathcal{R}}w\\ \bar{b} &= \bar{z}\frac{\partial z}{\partial b}\\ &= \bar{z}\end{aligned}$$

$$\mathbf{g}_i = \nabla\mathcal{L}(\mathbf{w},\mathbf{x}_i,t_i)$$

(GD)

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\sum_{i=1}^{N} g_i$$

(SGD)

$$i \sim \mathcal{U}[1,N], \quad \mathbf{w} \leftarrow \mathbf{w} - \eta g_i$$

(mSGD)

$$M \subset \{1,\ldots,N\}, \quad \mathbf{w} \leftarrow \mathbf{w} - \eta\sum_{i\in M}^{|M|} g_i$$

3 Misc

1. Activation functions $\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$ $\quad \sigma(z) = \frac{1}{1+\exp(-z)}$ $\quad \mathrm{ReLU}(z) = \max(0,z)$.
2. **Parametric** Benefits (i) Simpler (interpretability) (ii) Speed (iii) Less Data;
   Drawbacks (i) Constrained (ii) Limited Complexity (iii) Poor fit.
3. **Non-parametric** Benefits (i) Flexibility (ii) Power (No prior assumptions) (iii) Performance;
   Drawbacks (i) More data (ii) Slower (iii) Overfitting.
4. Decision of linear models: $\mathbf{W}\cdot\mathbf{x} + \mathbf{b} = \mathbf{0}$ (hyperplane).

SVM

0-1 loss: $\mathcal{L}_{0-1}(z,t) = \mathbb{I}\{\text{sign}(z) \neq t\}$

Hinge loss: $\mathcal{L}_{\text{H}}(z,t) = \max\{0, 1 - zt\}$

$$\min_{\mathbf{w},b} \Sigma_{i=1}^{N} \max\left\{0, 1 - t^{(i)} z^{(i)}(\mathbf{w}, b)\right\}$$

$$\min_{\mathbf{w},b} \Sigma_{i=1}^{N} \max\left\{0, 1 - t^{(i)} z^{(i)}(\mathbf{w}, b)\right\} + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

Optimize: gradient descent.

Boosting

Weighted training set: we can learn a classier using different costs (aka weights) for examples.

$$\Sigma_{n=1}^{N} w^{(n)} \mathbb{I}[h(x^{(n)}) \neq t^{(n)}]$$
$$w^(n) > 0 \wedge \Sigma_{n=1}^{N} w^{(n)} = 1$$

Decision Stump A decision tree with a single split.
AdaBoost **reduces bias** by making each classier focus on previous mistakes.