Supervised Learning :: CHEAT SHEET

Basic Concepts

Risk minimization

- Empirical risk minimizer:

$\hat{f} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right)$

- Optimal constant model: $\hat{f}_c = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} L(y^{(i)}, c)$
- Risk minimizer (Bayes optimal model):

$f^*_{\mathcal{H}_{all}} = \arg\min_{f \in \mathcal{H}_{all}} \mathcal{R}(f) = \arg\min_{f \in \mathcal{H}_{all}} \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right]$

The resulting risk is called **Bayes risk**: $\mathcal{R}^* = \mathcal{R}(f^*_{\mathcal{H}_{all}})$

- Bayes regret: $\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}^* = \underbrace{\left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}(f) - \mathcal{R}^*\right]}_{\text{approximation error}}$

Relative items

- **Residuals:** $r(\mathbf{x}) := y - f(\mathbf{x})$, best point-wise update $f(\mathbf{x}) \leftarrow f(\mathbf{x}) + r(\mathbf{x})$
- **Pseudo-residuals:** $\tilde{r}(\mathbf{x}) := -\frac{dL(y, f(\mathbf{x}))}{df(\mathbf{x})}$, approx. $f(\mathbf{x}) \leftarrow f(\mathbf{x}) + \tilde{r}(\mathbf{x})$
- **Margin:** $\nu(\mathbf{x}) := y \cdot f(\mathbf{x})$
- Prediction $\pi(\mathbf{x}) \in [0, 1]$ is called **calibrated** if

$$\mathbb{P}(y = 1 \mid \pi(\mathbf{x}) = p) = p \quad \forall p \in [0, 1]$$

- **Scoring rules** $S(Q, P) = \mathbb{E}_{y \sim Q}[L(Q, P)]$ is **proper** if true label distrib Q is among the optimal solutions, when we maximize $S(Q, P)$ in the 2nd argument (for a given Q)

$$S(Q, Q) \leq S(Q, P) \text{ for all } P, Q$$

Properties of Loss Functions

- Symmetric: $L\left(y, f(\mathbf{x})\right) = L(f(\mathbf{x}), y)$
- Translation-invariant: $L(y + a, f(\mathbf{x}) + a) = L\left(y, f(\mathbf{x})\right), a \in \mathbb{R}$
- Distance-based: can be written in terms of residual $L\left(y, f(\mathbf{x})\right) = \psi(r)$ for some $\psi : \mathbb{R} \to \mathbb{R}$, and $\psi(r) = 0 \Leftrightarrow r = 0$
- Robust: less influenced by outliers than by "inliers"

Properties of Optimization

- Smoothness: measured by number of continuous derivatives, depends on both $L\left(y, f(\mathbf{x})\right)$ and $f(\mathbf{x})$
- Convexity: have several good properties, depends on both $L\left(y, f(\mathbf{x})\right)$ and $f(\mathbf{x} \mid \boldsymbol{\theta})$

Regression Losses

- L2 Loss: convex, differentiable, sensitive to outliers, max. likelihood of Gaussian errors

$$L\left(y, f(\mathbf{x})\right) = (y - f(\mathbf{x}))^2 \quad \text{or} \quad L\left(y, f(\mathbf{x})\right) = 0.5(y - f(\mathbf{x}))^2$$

- L1 Loss: convex, more robust than L2, not differentiable at $y = f(\mathbf{x})$, not proper, max. likelihood of Laplace errors

$$L\left(y, f(\mathbf{x})\right) = |y - f(\mathbf{x})|$$

- Huber Loss: convex, once differentiable

$$L\left(y, f(\mathbf{x})\right) = \begin{cases} \frac{1}{2}(y - f(\mathbf{x}))^2 & \text{if } |y - f(\mathbf{x})| \leq \epsilon \\ \epsilon|y - f(\mathbf{x})| - \frac{1}{2}\epsilon^2 & \text{otherwise} \end{cases} \quad \epsilon > 0$$

- Log-cosh Loss: convex, twice differentiable

$$L\left(y, f(\mathbf{x})\right) = \log(\cosh(|y - f(\mathbf{x})|)) \qquad \cosh(x) = \frac{e^x + e^{-x}}{2}$$

- Cauchy Loss: differentiable, not convex

$$L\left(y, f(\mathbf{x})\right) = \frac{c^2}{2}\log\left(1 + \left(\frac{|y - f(\mathbf{x})|}{c}\right)^2\right), \quad c \in \mathbb{R}$$

- ϵ-Insensitive Loss: convex, not differentiable for $y - f(\mathbf{x}) \in \{-\epsilon, \epsilon\}$

$$L\left(y, f(\mathbf{x})\right) = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \leq \epsilon \\ |y - f(\mathbf{x})| - \epsilon & \text{otherwise} \end{cases}, \quad \epsilon \in \mathbb{R}_+$$

- Quantile Loss: extension of L1 with α-quantile as risk minimizer

$$L\left(y, f(\mathbf{x})\right) = \begin{cases} (1 - \alpha)(f(\mathbf{x}) - y) & \text{if } y < f(\mathbf{x}) \\ \alpha(y - f(\mathbf{x})) & \text{if } y \geq f(\mathbf{x}) \end{cases}, \quad \alpha \in (0, 1)$$

Classification Losses

- 0-1 Loss: not continuous, NP hard, proper but not strict

h discrete classifier, f score function, π probability function

$$L\left(y, h(\mathbf{x})\right) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}, \mathcal{R}^* = 1 - \mathbb{E}_x[\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x})]$$

$$L\left(y, f(\mathbf{x})\right) = \mathbb{1}_{\{\nu < 0\}} = \mathbb{1}_{\{yf(\mathbf{x}) < 0\}} \qquad y \in \{-1, +1\}$$

$$L\left(y, \pi(\mathbf{x})\right) = y\mathbb{1}_{\{\pi(\mathbf{x}) < 0.5\}} + (1 - y)\mathbb{1}_{\{\pi(\mathbf{x}) \geq 0.5\}} = \mathbb{1}_{\{(2y-1)(\pi(\mathbf{x}) - 0.5) < 0\}} \; y \in \{0, 1\}$$

- Bernoulli Loss: strictly proper, max. likelihood of Bernoulli errors

$$L\left(y, \pi(\mathbf{x})\right) = -y\log(\pi(\mathbf{x})) - (1 - y)\log(1 - \pi(\mathbf{x})) \qquad y \in \{0, 1\}$$

$$L\left(y, \pi(\mathbf{x})\right) = -\frac{1 + y}{2}\log(\pi(\mathbf{x})) - \frac{1 - y}{2}\log(1 - \pi(\mathbf{x})) \qquad y \in \{-1, +1\}$$

$$L\left(y, f(\mathbf{x})\right) = -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))) \qquad y \in \{0, 1\}$$

$$L\left(y, f(\mathbf{x})\right) = \log(1 + \exp(-y \cdot f(\mathbf{x}))) \qquad y \in \{-1, +1\}$$

- Brier Score: strictly proper

$$L\left(y, \pi(\mathbf{x})\right) = (\pi(\mathbf{x}) - y)^2, y \in \{0, 1\}$$

$$L\left(y, f(\mathbf{x})\right) = ((1 + \exp(-f(\mathbf{x})))^{-1} - y)^2, y \in \{0, 1\}$$

- Hinge Loss: continuous, convex, upper bound on 0-1-loss

$$L\left(y, f(\mathbf{x})\right) = \max\{0, 1 - yf(\mathbf{x})\} \qquad y \in \{-1, +1\}$$

- Squared Hinge Loss: continuous, convex, more outlier-sensitive than hinge loss

$$L\left(y, f(\mathbf{x})\right) = \max\{0, (1 - yf(\mathbf{x}))\}^2 \qquad y \in \{-1, +1\}$$

- Exponential Loss: convex, differentiable

$$L\left(y, f(\mathbf{x})\right) = \exp(-yf(\mathbf{x})) \qquad y \in \{-1, +1\}$$

- AUC-Loss: not differentiable

$$AUC = \frac{1}{n_+}\frac{1}{n_-}\sum_{i: y^{(i)} = 1}\sum_{j: y^{(j)} = -1} \mathbb{I}[f^{(i)} > f^{(j)}]$$

$y \in \{-1, +1\}$ with n_- negative and n_+ positive samples

- Multiclass Bernoulli Loss

$$L\left(y, \pi(\mathbf{x})\right) = -\sum_{k=1}^{g}[y = k]\log(\pi_k(\mathbf{x}))$$

Risk minimization is equivalent to **entropy splitting**

Entropy of node $\mathcal{N}$: $\text{Imp}(\mathcal{N}) = -\sum_{k=1}^{g}\pi_k^{(\mathcal{N})}\log\pi_k^{(\mathcal{N})}$

- Multiclass Brier Score: strictly proper

$$L\left(y, \pi(\mathbf{x})\right) = \sum_{k=1}^{g}([y = k] - \pi_k(\mathbf{x}))^2$$

Risk minimization is equivalent to **Gini splitting**

Gini index of node $\mathcal{N}$: $\text{Imp}(\mathcal{N}) = \sum_{k=1}^{g}\pi_k^{(\mathcal{N})}(1 - \pi_k^{(\mathcal{N})})$

Supervised Learning :: CHEAT SHEET

Logistic Regression

Given n observations $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{0, 1\}$, we want to minimize:

$$\mathcal{R}_{\text{emp}} = -\sum_{i=1}^{n} y^{(i)} \log(\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)) + (1 - y^{(i)} \log(1 - \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)))$$

Probabilistic classifier: $\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = s(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))$
Sigmoid function: $s(f) = \frac{1}{1+\exp(-f)}$, $\frac{\partial}{\partial f} s(f) = s(f)(1 - s(f))$

Score: $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \mathbf{x}$., $\frac{\partial f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = (\mathbf{x}^{(i)})^\top$.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \sum_{i=1}^{n} (\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) - y^{(i)})(\mathbf{x}^{(i)})^\top = (\pi(\mathbf{X} | \boldsymbol{\theta}) - \mathbf{y})^\top \mathbf{X}$$

where $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^\top \in \mathbb{R}^{n \times d}$, $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^\top$,
$\pi(\mathbf{X} | \boldsymbol{\theta}) = \left(\pi\left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta}\right), \ldots, \pi\left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta}\right)\right)^\top \in \mathbb{R}^n$.

$$\nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \sum_{i=1}^{n} \mathbf{x}^{(i)} (\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) (1 - \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)))(\mathbf{x}^{(i)})^\top = \mathbf{X}^\top \mathbf{D} \mathbf{X}$$

Bias-Variance Decomposition

$$GE_n(\mathcal{I}) = \underbrace{\sigma^2}_{\text{Var. of } \epsilon} + \mathbb{E}_x \underbrace{\left[\text{Var}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x})\right]}_{\text{Variance of learner at } \mathbf{x}} + \mathbb{E}_x \underbrace{\left[(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x})))^2 \mid \mathbf{x}\right]}_{\text{Squared bias of learner at } \mathbf{x}}$$

1. First: variance of "pure" **noise** ϵ; aka Bayes, intrinsic or irreducible error; whatever we we do, will never be better
2. Second: how much $\hat{f}_{\mathcal{D}_n}(\mathbf{x})$ **fluctuates** at test $\mathbf{x}$ if we vary training data, averaged over feature space; = learner's tendency to learn random things irrespective of real signal (overfitting)
3. Third: how "off" are we on average at test locations (underfitting); uses "average model integrated out over all $\mathcal{D}_n$"; models with high capacity have low **bias** and vice versa

Summary of Loss Functions and Estimators

Loss Function	Risk Minimizer	Optimal Constant Model
L2	$f^*(\mathbf{x}) = \mathbb{E}_{y\mid x}[y \mid \mathbf{x}]$	$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
L1	$f^*(\mathbf{x}) = \text{med}_{y\mid x}[y \mid \mathbf{x}]$	$f(\mathbf{x}) = \text{med}(y^{(i)})$
0-1	$h^*(\mathbf{x}) = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x})$	$h(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}$
Brier	$\pi_k^*(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x})$	$\pi_k(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)}=k\}}$
Bernoulli (on probs)	$\pi_k^*(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x})$	$\pi_k(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)}=k\}}$
Bernoulli (on scores)	$f_k^*(\mathbf{x}) = \log\left(\frac{\mathbb{P}(y=k \mid \mathbf{x})}{1-\mathbb{P}(y=k \mid \mathbf{x})}\right)$	$f_k(\mathbf{x}) = \log \frac{n_k}{n - n_k}$