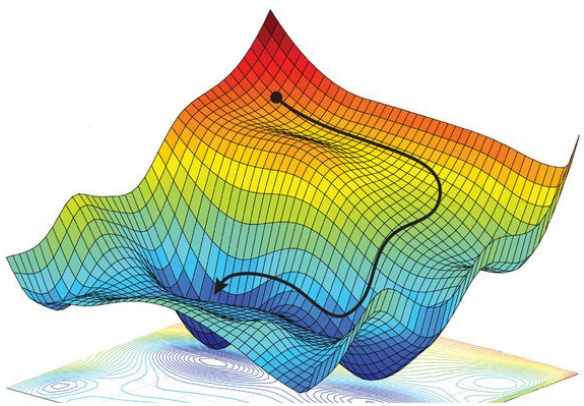
Introduction to Machine Learning

Theoretical Considerations on Classification Losses



Learning goals

- Understand what the Bayes optimal model is
- Derive the point-wise optimizers of different loss functions
- Derivce the optimal constant models

BAYES OPTIMAL MODEL

POINT-WISE OPTIMUM

We can in general rewrite the risk as

$$\begin{aligned} \mathcal{R}(f) &= \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}[L(y, f(\mathbf{x}))]\right] \\ &= \mathbb{E}_x\left[\sum_{k \in \mathcal{Y}} L(k, f(\mathbf{x}))\mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x})\right], \end{aligned}$$

with $\mathbb{P}(y = k|\mathbf{x} = \mathbf{x})$ being the posterior probability for class k.

The optimal model for a loss function $L\left(y, f(\mathbf{x})\right)$ is

$$\hat{f}(\mathbf{x}) = \arg\min_{f \in \mathcal{H}} \sum_{k \in \mathcal{Y}} L(k, f(\boldsymbol{x}))\mathbb{P}(y = k|\mathbf{x} = \mathbf{x}).$$

POINT-WISE OPTIMUM

If we can estimate $\mathbb{P}_{xy}$ very well via $\pi_k(\mathbf{x})$ through a stochastic model, we can now compute the loss-optimal classifications point-wise.
But usually we directly adapt to the loss via **empirical risk minimization**.

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

0-1-Loss

0-1-LOSS: POINT-WISE OPTIMUM

For an (unrestricted) classifier $h(\mathbf{x})$ and the 0-1-loss:

$$\min_{h \in \mathcal{H}} \mathcal{R}(h) = \mathbb{E}_{xy}[L(y, h(\mathbf{x}))].$$

The (point-wise) solution of the above minimization problem is

$$\begin{aligned}
\hat{h}(\mathbf{x}) &= \arg\min_{l \in \mathcal{Y}} \sum_{k \in \mathcal{Y}} L(k, l) \cdot \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\min_{l \in \mathcal{Y}} \sum_{k \neq l} \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) = \arg\min_{l \in \mathcal{Y}} 1 - \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x})
\end{aligned}$$

which corresponds to predicting the most probable class.

$\hat{h}(\mathbf{x})$ is called the **Bayes optimal classifier**. The expected loss is called **Bayes loss** or **Bayes error rate** for the 0-1-loss.

0-1-LOSS: OPTIMAL CONSTANT MODEL

The optimal constant model (featureless predictor) under 0-1 loss, with $y \in \{-1, +1\}$, either for hard classifiers $h(\mathbf{x})$ or scoring classifiers $f(\mathbf{x})$

$$L(y, h(\mathbf{x})) = \mathbb{1}_{y \neq h(\mathbf{x})}$$

is the classifier that predicts the most frequent class in the data

$$h(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\} \qquad \text{or} \qquad f(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}.$$

Proof: Exercise / Trivial.

BRIER SCORE: POINT-WISE OPTIMUM

The minimizer of the (theoretical) risk $\mathcal{R}(f)$ for the Brier score

$$\hat{\pi}(\mathbf{x}) \quad = \quad \mathbb{P}(y \mid \mathbf{x} = \mathbf{x}),$$

which means that the Brier score would reach its minimum if the prediction equals the "true" probability of the outcome.

Proof: We have seen that the (theoretical) optimal prediction c for an arbitrary loss function at fixed point $\mathbf{x}$ is

$$\arg\min_{c} \sum_{k \in \mathcal{Y}} L(y, c) \mathbb{P}(y = k | \mathbf{x} = \mathbf{x}) \, .$$

BRIER SCORE: POINT-WISE OPTIMUM

We plug in the Brier score

$$\begin{aligned} & \arg\min_c L(1,c) \underbrace{\mathbb{P}(y=1|\mathbf{x}=\mathbf{x})}_{p} + L(0,c) \underbrace{\mathbb{P}(y=0|\mathbf{x}=\mathbf{x})}_{1-p} \\ = \quad & \arg\min_c \quad (c-1)^2 p + c^2(1-p) \\ = \quad & \arg\min_c \quad (c-p)^2. \end{aligned}$$

The expression is minimal if $c = p = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$.

BRIER SCORE: OPTIMAL CONSTANT MODEL

The optimal constant probability model $\pi(\mathbf{x}) = \theta$ w.r.t. the Brier score for labels from $\mathcal{Y} = \{0, 1\}$ is:

$$
\begin{aligned}
\min_{\theta} \mathcal{R}_{\text{emp}}(\theta) &= \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \theta \right)^2 \\
\Leftrightarrow \frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} &= -2 \cdot \sum_{i=1}^{n} (y^{(i)} - \theta) = 0 \\
\hat{\theta} &= \frac{1}{n} \sum_{i=1}^{n} y^{(i)}.
\end{aligned}
$$

This is the fraction of class-1 observations in the observed data.
(This also directly follows from our *L2*-proof for regression).

BERNOULLI LOSS: POINT-WISE OPTIMUM

The theoretical point-wise optimum for scores under Bernoulli loss is actually the point-wise log-odds:

$$\hat{f}(\mathbf{x}) \quad = \quad \ln\left(\frac{\mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}{1 - \mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}\right).$$

The function is undefined when $P(y \mid \mathbf{x} = \mathbf{x}) = 1$ or $P(y \mid \mathbf{x} = \mathbf{x}) = 0$, but predicts a smooth curve which grows when $P(y \mid \mathbf{x} = \mathbf{x})$ increases and equals 0 when $P(y \mid \mathbf{x} = \mathbf{x}) = 0.5$.

Proof: We consider the case $\mathcal{Y} = \{-1, 1\}$. We have seen that the (theoretical) optimal prediction c for an arbitrary loss function at fixed point $\mathbf{x}$ is

$$\arg\min_{c} \sum_{k \in \mathcal{Y}} L(y, c) \mathbb{P}(y = k | \mathbf{x} = \mathbf{x}) \, .$$

BERNOULLI LOSS: POINT-WISE OPTIMUM

We plug in the Bernoulli loss

$$\begin{aligned} & \arg\min_c L(1,c) \underbrace{\mathbb{P}(y=1|\mathbf{x}=\mathbf{x})}_{p} + L(-1,c) \underbrace{\mathbb{P}(y=-1|\mathbf{x}=\mathbf{x})}_{1-p} \\ = \quad & \arg\min_c \ln(1+\exp(-c))p + \ln(1+\exp(c))(1-p). \end{aligned}$$

Setting the derivative w.r.t. c to zero yields

$$\begin{aligned} 0 &= -\frac{\exp(-c)}{1+\exp(-c)}p + \frac{\exp(c)}{1+\exp(c)}(1-p) \\ &= -\frac{\exp(-c)}{1+\exp(-c)}p + \frac{1}{1+\exp(-c)}(1-p) \\ &= -p + \frac{1}{1+\exp(-c)} \\ \Leftrightarrow p &= \frac{1}{1+\exp(-c)} \\ \Leftrightarrow c &= \ln\left(\frac{p}{1-p}\right) \end{aligned}$$

BERNOULLI: OPTIMAL CONSTANT MODEL

The optimal constant probability model $\pi(\mathbf{x}) = \theta$ w.r.t. the Bernoulli loss for labels from $\mathcal{Y} = \{0, 1\}$) is:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

Again, this is the fraction of class-1 observations in the observed data. We can simply prove this again by setting the derivative of the risk to 0 and solving for θ.

BERNOULLI: OPTIMAL CONSTANT MODEL

The optimal constant score model $f(\mathbf{x}) = \theta$ w.r.t. the Bernoulli loss labels from $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$ is:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \ln \frac{n_{+1}}{n_{-1}} = \ln \frac{n_{+1}/n}{n_{-1}/n}$$

where n_{-1} and n_{+1} are the numbers of negative and positive observations, respectively.

This again shows a tight (and unsurprising) connection of this loss to log-odds.

Proving this is also a (quite simple) exercise.

BERNOULLI-LOSS: NAMING CONVENTION

We have seen three loss functions that are closely related. In the literature, there are different names for the losses:

$$\begin{aligned} L_{-1+1}(y, f(\mathbf{x})) &= \ln(1 + \exp(-yf(\mathbf{x}))) \\ L_{0,1}(y, f(\mathbf{x})) &= -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))). \end{aligned}$$

are referred to as Bernoulli, Binomial or logistic loss.

$$L_{0,1}(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})).$$

is referred to as cross-entropy or log-loss.

For simplicity, we will call all of them **Bernoulli loss**, and rather make clear whether they are defined on labels $y \in \{0, 1\}$ or $y \in \{-1, 1\}$ and on scores $f(\mathbf{x})$ or probabilities $\pi(\mathbf{x})$.