Parametrized probabilistic classifiers

- probabilistic classifier G_θ depends on parameter θ
- we'll choose θ by ERM or RERM
- we judge probabilistic classifier by average negative log likelihood on a test data set

ERM for probabilistic classifiers

- data set $u^1, \ldots, u^n$, $v^1, \ldots, v^n$
- parametrized probabilistic classifier G_θ, with predicted distributions $\hat{p}^1, \ldots, \hat{p}^n$ (which depend on θ)
- define a loss function $\ell(\hat{p}, v)$
  - first argument $\hat{p}$ is a *distribution* on $\mathcal{V}$
  - second argument v is an *element* of $\mathcal{V}$
- ERM: choose θ to minimize the average loss $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{p}^i, v^i)$
- RERM: choose θ to minimize the average loss plus a regularizer, $\mathcal{L}(\theta) + \lambda r(\theta)$
- $\lambda \geq 0$ is the regularization hyper-parameter

Cross-entropy loss

Negative log likelihood

- the *negative log-likelihood* of v under distribution $\hat{p}$ is

$$\ell^{\text{ce}}(\hat{p}, v) = -\log \hat{p}(v)$$

  i.e., the negative log of the probability of the outcome v

- ℓ^{ce} takes two arguments, the first is a function p, the second is an element of $\mathcal{V}$
- since $\hat{p}(v) \leq 1$, $\ell^{\text{ce}}(\hat{p}, v) \geq 0$
- $\ell^{\text{ce}}(\hat{p}, v) = 0$ only if $\hat{p}(v) = 1$, *i.e.*, we are certain about the outcome and we're right
- we want the negative log-likelihood to be small

Cross-entropy loss

- $\ell^{\mathrm{ce}}(\hat{p}, v)$ is a *loss function* for probabilistic prediction
  - similarly to loss function $\ell(\hat{y}, y)$ for deterministic predictions, it compares the predicted value $\hat{y}$ with the actual value y
  - but it takes a predicted probability $\hat{p}$ instead of a point prediction $\hat{y}$
  - and it takes a raw target v instead of an embedded target $y = \psi(v)$
- using this, we can compute the *empirical risk* on a data set $u^1, \dots, u^n$, $v^1, \dots, v^n$

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \ell^{\mathrm{ce}}(\hat{p}^i, v^i) = \frac{1}{n} \sum_{i=1}^{n} \ell^{\mathrm{ce}}(G(u^i), v^i)$$

- the empirical risk is the *average negative log likelihood* which we'd like to be small
- ℓ^{ce} is called the *cross-entropy loss*
- average cross-entropy loss is the cross entropy, when $\hat{p}$ is constant

Cross-entropy loss

- $\ell^{\text{ce}}(\hat{p}, v)$ is how implausible v with distribution $\hat{p}$
  - ℓ^{ce} small means v is 'typical'
  - ℓ^{ce} large means v is 'unlikely'
- other names for ℓ^{ce}: surprise, perplexity, ...

Logistic un-embedding

Un-embedding for probabilistic classification

- in point classification, we un-embed $\hat{y} \in \mathbf{R}^K$ as $\hat{v} = v_i$, with $i = \operatorname{argmin}_j ||\hat{y} - \psi_j||_2$
- this un-embedding maps $\mathbf{R}^k$ into $\mathcal{V} = \{v_1, \ldots, v_K\}$

- for probabilistic classification, we un-embed $\hat{y} \in \mathbf{R}^K$ as $\hat{p} = \sigma(\hat{y})$, the distribution on $\mathcal{V}$ given by

$$\hat{p}(v_k) = \frac{\exp \hat{y}_k}{\sum_{j=1}^K \exp \hat{y}_j}, \quad k = 1, \ldots, K$$

- σ is called the *logistic map*, *activation function*, *inverse link function*, *softargmax function*, *normalized exponential* or *softmax function*
- this un-embedding maps a vector $\hat{y} \in \mathbf{R}^K$ to a probability distribution on $\mathcal{V}$

Properties of logistic map

$$\hat{p}(v_k) = \frac{\exp \hat{y}_k}{\sum_{j=1}^{K} \exp \hat{y}_j}, \quad k = 1, \dots, K$$

- adding constant to each entry of $\hat{y}$ doesn't affect $\hat{p}$
- increasing $\hat{y}_k$ (leaving over entries the same) increases $\hat{p}(v_k)$, decreases $\hat{p}(v_j)$ for $j \neq k$
- $\hat{p}(v_k)$ can be close to, but not equal to, zero or one
- $\hat{p}(v_k)$ is close to zero or one when $\hat{y}_k$ is very much less than, or greater than, the other entries
- if $\hat{y} = 0$ (or all its entries are equal), $\hat{p}(v_k) = 1/K$ for all k, so is $\hat{p}$ is the uniform distribution

ERM with logistic un-embedding

ERM with logistic un-embedding

- for deterministic classification, we embed $x^i = \phi(u^i)$, $y^i = \psi(v^i)$, and ERM minimizes

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \ell(g_\theta(x^i), y^i)$$

  resulting predictor is $\hat{v} = \psi^\dagger(g_\theta(\phi(x)))$

- for probabilistic classification, we embed $x^i = \phi(u^i)$, and ERM minimizes

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \ell^{\text{ce}}(\sigma(g_\theta(x^i)), v^i)$$

  resulting predictor is $\hat{p} = \sigma(g_\theta(\phi(x)))$

Logistic loss

- assume our probability guesses $\hat{p}$ comes from logistic un-embedding, $\hat{p} = \sigma(\hat{y})$
- what is the cross-entropy loss of our guess $\hat{p}$, when true label is $v = v_k$?

$$\ell^{\text{ce}}(\hat{p}, v_k) = \ell^{\text{ce}}(\sigma(\hat{y}), v_k) = -\log\left(\frac{\exp \hat{y}_k}{\sum_{j=1}^{K} \exp \hat{y}_j}\right) = -\hat{y}_k + \log\left(\sum_{j=1}^{K} \exp \hat{y}_j\right)$$

- this is the logistic loss (with $\kappa_i = 1$)

$$\ell(\hat{y}, \psi_k) = -\hat{y}_i + \log\left(\sum_{j=1}^{K} \exp \hat{y}_j\right)$$

- so with logistic un-embedding, *logistic loss* is the cross-entropy loss, and the resulting empirical risk is the *average negative log-likelihood*

Interpreting multi-class logistic regression

- logistic regression yields probabilistic classifier

$$\hat{p} = \sigma(\hat{y}) = \frac{\exp \hat{y}}{\mathbf{1}^\mathsf{T} \exp \hat{y}}, \qquad \hat{y} = \theta^\mathsf{T} x$$

- assume $x_1 = 1$ is constant feature, other features standardized
- first row of θ, θ_1^T, is $\hat{y}$ when $x_{2:d} = 0$, *i.e.*, all non-constant features take their mean value (zero)
- corresponding distribution is $\hat{p} = \sigma(\theta_1)$
- θ_{ij} gives effect of x_i on $\hat{p}_j$

Logistic un-embedding for Boolean classification

Boolean probabilistic classifier

- Boolean case: $\mathcal{V} = \{v_1, v_2\}$
- given u, we guess $\hat{p} = G(u)$
- to specify the function $\hat{p}$, we have to give the two numbers $\hat{p}(v_1)$ and $\hat{p}(v_2)$
- we can just give one of them, since they sum to one
- *e.g.*, we can give the number $\hat{p}(v_2)$, the probability that $v = v_2$; we have $\hat{p}(v_1) = 1 - \hat{p}(v_2)$

- example: to predict probability of rain or shine, we can give just $\hat{p}(\textsc{rain})$, since $\hat{p}(\textsc{shine}) = 1 - \hat{p}(\textsc{rain})$

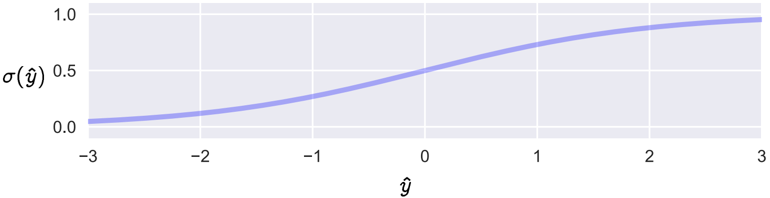
Un-embedding for Boolean probabistic classification

- the function $\sigma(\hat{y}) = \dfrac{1}{1+e^{-\hat{y}}}$ is called the *sigmoid* function
- we use σ for both the sigmoid and the logistic functions, since both are activation functions mapping $\mathbf{R}^m$ to probability distributions on $\mathcal{V}$
- in the Boolean case, can use $\hat{y} \in \mathbf{R}$ instead of $\hat{y} \in \mathbf{R}^2$
- when $\mathcal{V} = \{v_1, v_2\}$, we can un-embed via

$$\hat{p}(v_1) = \sigma(\hat{y}) \qquad \hat{p}(v_2) = 1 - \sigma(\hat{y})$$

- maps $\hat{y} \in \mathbf{R}$ to a distribution on $\mathcal{V}$
- the inverse function $\hat{y} = \log \dfrac{\hat{p}(v_1)}{1-\hat{p}(v_1)}$ is called the *log-odds* or *logit* function

Sigmoid function

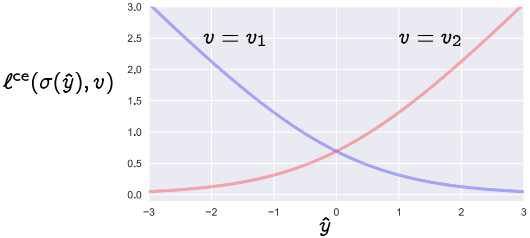


- the sigmoid function $\sigma(\hat{y}) = \dfrac{1}{1 + e^{-\hat{y}}}$
- has symmetry property $\sigma(-\hat{y}) = 1 - \sigma(\hat{y})$

Boolean logistic loss

- using the sigmoid un-embedding, we have

$$
\begin{aligned}
\ell^{\text{ce}}(\hat{p}^i, v^i) &= \ell^{\text{ce}}(\sigma(\hat{y}^i), v^i) \\
&= \begin{cases} -\log \sigma(\hat{y}^i) & \text{if } v^i = v^1 \\ -\log(1 - \sigma(\hat{y}^i)) & \text{if } v^i = v^2 \end{cases} \\
&= \begin{cases} \log(1 + e^{-\hat{y}^i}) & \text{if } v^i = v^1 \\ \log(1 + e^{\hat{y}^i}) & \text{if } v^i = v^2 \end{cases} \\
&= \begin{cases} \ell(\hat{y}, 1) & \text{if } v^i = v^1 \\ \ell(\hat{y}, -1) & \text{if } v^i = v^2 \end{cases}
\end{aligned}
$$



- so with this un-embedding the cross-entropy loss is the Boolean logistic loss

Empirical risk minimization

- empirical risk is

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \ell^{\mathrm{ce}}(\hat{p}^i, v^i) = \frac{1}{n}\left(\sum_{i:v^i=v_1} -\log(\sigma(\theta^\mathsf{T} x)) + \sum_{i:v^i=v_2} -\log(\sigma(-\theta^\mathsf{T} x))\right)$$

- choose θ to minimize empirical risk
- then $\sigma(\theta^\mathsf{T} x)$ is the predicted probability that $v = v_1$ at x

Empirical risk minimization

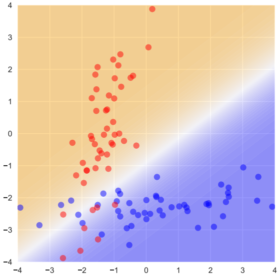
- empirical risk is

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \ell^{\mathsf{ce}} = \frac{1}{n}\sum_{i=1}^{n} -\log(\sigma(\theta^\mathsf{T} x^i)_{y^i})$$

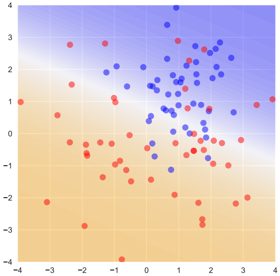
- choose θ to minimize empirical risk
- then $\sigma(\theta^\mathsf{T} x)$ is the predicted probability distribution at x

Examples

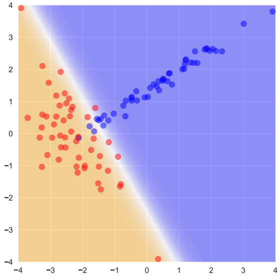
Examples



$||\theta|| \approx 2.5$



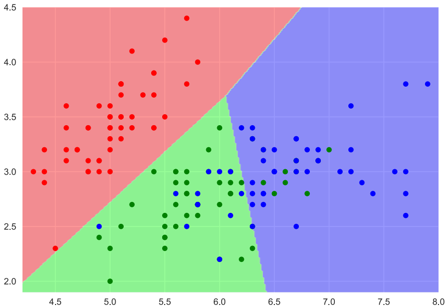
$||\theta|| \approx 1.25$

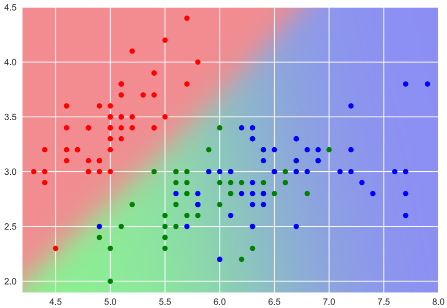


$||\theta|| \approx 8.1$

- larger θ corresponds to greater certainty

Examples





- left-hand plot shows which component of $\theta^{\mathsf{T}} x$ is largest
- right-hand plot shows $\sigma(\theta^{\mathsf{T}} x)$

Summary

Summary

- we judge a probabilistic classifier by its average log likelihood on test data
- this equals the empirical risk, when using the cross-entropy loss
- we un-embed a prediction $\hat{y} \in \mathbf{R}^K$ into a distribution using the logistic un-embedding