18-661 Introduction to Machine Learning

Multi-class Classification

Fall 2020

ECE – Carnegie Mellon University

Announcements

- HW2 is due on Friday.
- HW3 will be released by Friday.
- HW1 grades have been released on Gradescope. Regrade requests will remain open for a week. The solution of HW1 is also posted.

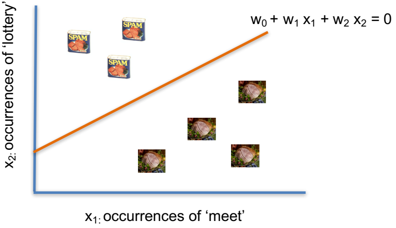
Outline

1. Review of Logistic Regression
2. Non-linear Decision Boundaries
3. Multi-class Classification
   - Multi-class Naive Bayes
   - Multi-class Logistic Regression

Review of Logistic Regression

Intuition: Logistic regression

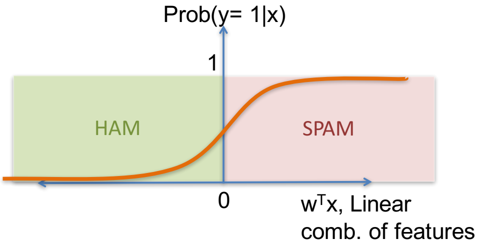
- x_1 = # of times 'meet' appears in an email
- x_2 = # of times 'lottery' appears in an email
- Define feature vector $\mathbf{x} = [1, x_1, x_2]$
- Learn the decision boundary $w_0 + w_1 x_1 + w_2 x_2 = 0$ such that
  - If $\mathbf{w}^\top \mathbf{x} \geq 0$ declare $y = 1$ (spam)
  - If $\mathbf{w}^\top \mathbf{x} < 0$ declare $y = 0$ (ham)



A linear classifier maps features into points in a high-dimensional space, and uses hyperplanes to separate them into classes.

Intuition: Logistic regression

- Suppose we want to output the probability of an email being spam/ham instead of just 0 or 1...
- This gives information about the confidence in the decision!
- Use a function $\sigma(\mathbf{w}^\top \mathbf{x})$ that maps $\mathbf{w}^\top \mathbf{x}$ to a value between 0 and 1.



Probability that predicted label is 1 (spam)

Our goal: Finding optimal weights $\mathbf{w}$ that accurately predict this probability for a new email.

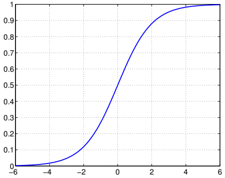
Formal setup: Binary logistic classification

- Input: $\boldsymbol{x} = [1, x_1, x_2, \ldots x_D] \in \mathbb{R}^{D+1}$
- Output: $y \in \{0, 1\}$
- Training data: $\mathcal{D} = \{(\boldsymbol{x}_n, y_n), n = 1, 2, \ldots, N\}$
- Model:

$$P(y = 1|\boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$$

and $\sigma[\cdot]$ stands for the *sigmoid* function

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$



How to optimize w?

- Probability of a single training sample $(\boldsymbol{x}_n, y_n)$

$$P(y_n|\boldsymbol{x}_n; \boldsymbol{w}) = \begin{cases} \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) & \text{if} \quad y_n = 1 \\ 1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) & \text{otherwise} \end{cases}$$

- Compact expression, exploiting that y_n is either 1 or 0

$$P(y_n|\boldsymbol{x}_n; \boldsymbol{w}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)^{y_n}[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)]^{1-y_n}$$

- Minimize the negative log-likelihood of the whole training data $\mathcal{D}$, i.e. the cross-entropy error function

$$\mathcal{E}(\boldsymbol{w}) = -\sum_n \{y_n \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) + (1 - y_n) \log[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)]\}$$

Gradient descent for logistic regression

- We want to minimize the cross-entropy error function:

$$\mathcal{E}(\boldsymbol{w}) = -\sum_n \{y_n \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) + (1 - y_n)\log[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)]\}$$

- Simple fact: derivatives of $\sigma(a)$ have a nice form:

$$\frac{d}{d\,a}\sigma(a) = \sigma(a)[1 - \sigma(a)]$$

- Gradient of cross-entropy loss is then

$$\frac{\partial \mathcal{E}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \sum_n \underbrace{\left\{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) - y_n\right\}}_{:=e_n} \boldsymbol{x}_n$$

- $e_n = \left\{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) - y_n\right\}$ is called the *error* for the nth training sample.

Gradient descent for logistic regression

- Choose a proper step size $\eta > 0$.
- Iteratively update the parameters following the negative gradient to minimize the error function

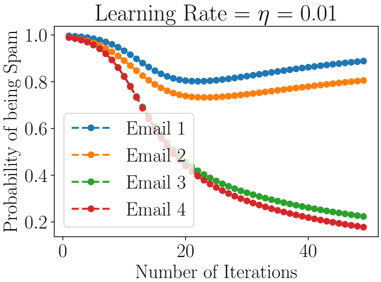
$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta \sum_n \left\{ \sigma(\boldsymbol{x}_n^\top \boldsymbol{w}^{(t)}) - y_n \right\} \boldsymbol{x}_n.$$

- Can instead perform stochastic gradient descent by randomly choosing a data point (with a possibly different learning rate η)

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta \left\{ \sigma(\boldsymbol{x}_{i_t}^\top \boldsymbol{w}^{(t)}) - y_{i_t} \right\} \boldsymbol{x}_{i_t}$$

where i_t is drawn uniformly at randomly from the training data $\{1, 2, \cdots\}$.

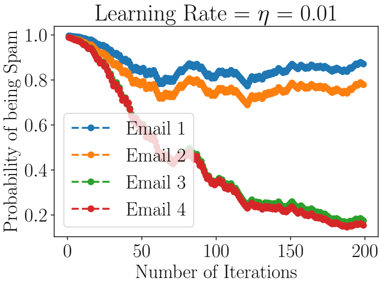
Batch gradient descent vs. SGD



Batch GD

fewer iterations,

every iteration uses all samples



SGD

more iterations,

every iteration uses one sample

Logistic regression vs. linear regression

	Logistic regression	**Linear regression**
Training data	$(\boldsymbol{x}_n, y_n), y_n \in \{0,1\}$	$(\boldsymbol{x}_n, y_n), y_n \in \mathbb{R}$
Loss function	cross-entropy	RSS
Interpretation of $y_n \mid \boldsymbol{x}_n, \boldsymbol{w}$	$\sim \mathrm{Ber}(\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))$	$\sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2)$
Gradient per sample	$\left(\sigma(\boldsymbol{x}_n^\top \boldsymbol{w}) - y_n\right) \boldsymbol{x}_n$	$\left(\boldsymbol{x}_n^\top \boldsymbol{w} - y_n\right) \boldsymbol{x}_n$

Cross-entropy loss function (logistic regression):

$$\mathcal{E}(\boldsymbol{w}) = -\sum_n \{y_n \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) + (1 - y_n) \log[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)]\}$$

RSS loss function (linear regression):

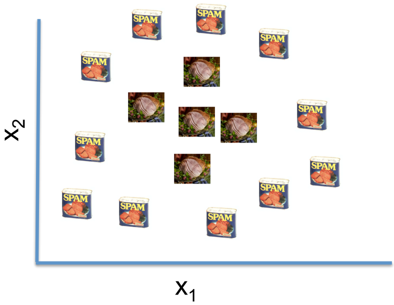
$$RSS(\boldsymbol{w}) = \frac{1}{2} \sum_n (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

Outline

1. Review of Logistic Regression
2. Non-linear Decision Boundaries
3. Multi-class Classification
   - Multi-class Naive Bayes
   - Multi-class Logistic Regression

Non-linear Decision Boundaries

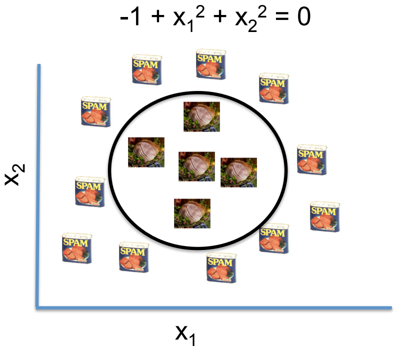
How to handle more complex decision boundaries?



- This data is not linearly separable...
- Use non-linear basis functions to add more features.

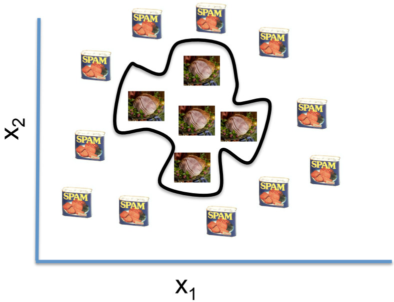
Adding polynomial features

- New feature vector is $\mathbf{x} = [1, x_1, x_2, x_1^2, x_2^2]$
- $\Pr(y = 1|\mathbf{x}) = \sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2)$
- If $\mathbf{w} = [-1, 0, 0, 1, 1]$, the boundary is $-1 + x_1^2 + x_2^2 = 0$
  - If $-1 + x_1^2 + x_2^2 \geq 0$ declare spam
  - If $-1 + x_1^2 + x_2^2 < 0$ declare ham



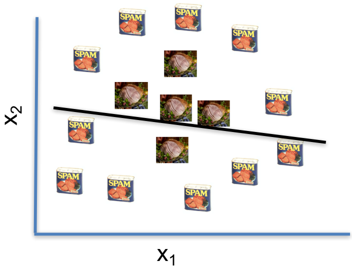
Adding polynomial features

- What if we add many more features and define $\mathbf{x} = [1, x_1, x_2, x_1^2, x_2^2, x_1^3, x_2^3, \ldots]$?
- We get a complex decision boundary

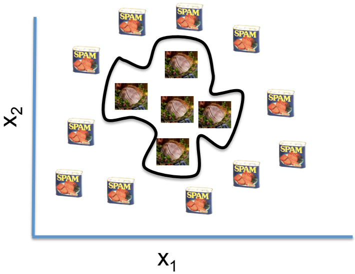


Can result in overfitting and bad generalization to new data points.

Concept-check: Bias-Variance Trade-off



high bias



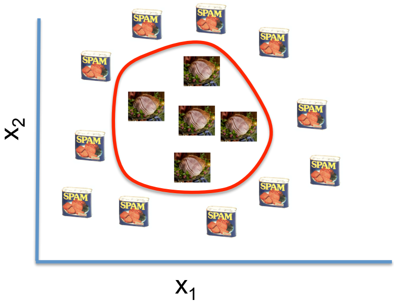
high variance

Solution to overfitting: Regularization

- Add regularization term to be cross entropy loss function

$$\mathcal{E}(\boldsymbol{w}) = -\sum_n \{y_n \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) + (1-y_n) \log[1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)]\} + \underbrace{\frac{1}{2}\lambda\|w\|_2^2}_{\text{regularization}}$$

- Perform gradient descent on this regularized function
- Often, we do NOT regularize the bias term w_0



Outline

1. Review of Logistic Regression

2. Non-linear Decision Boundaries

3. Multi-class Classification

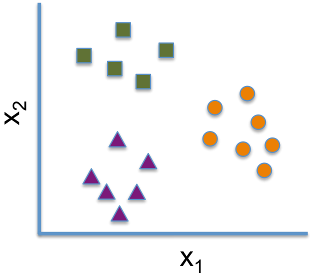
    Multi-class Naive Bayes

    Multi-class Logistic Regression

Multi-class Classification

What if there are more than 2 classes?

- Dog vs. cat. vs crocodile
- Movie genres (action, horror, comedy, . . .)
- Part of speech tagging (verb, noun, adjective, . . .)
- . . .



Setup

Predict multiple classes/outcomes $C_1, C_2, \ldots, C_M$:

- Weather prediction: sunny, cloudy, raining, etc
- Optical character recognition: 10 digits + 26 characters (lower and upper cases) + special characters, etc.

M =number of classes

Methods we've studied for binary classification:

- Naive Bayes
- Logistic regression

Do they generalize to multi-class classification?

Naive Bayes is already multi-class!

Formal Definition

Given a random vector $\mathbf{X} \in \mathbb{R}^{\mathrm{K}}$ and a dependent variable $Y \in [\mathrm{C}]$, the Naive Bayes model defines the joint distribution

$$
\begin{aligned}
P(\mathbf{X} = \mathbf{x}, Y = c) &= P(Y = c)P(\mathbf{X} = \mathbf{x}|Y = c) && (1)\\
&= P(Y = c)\prod_{k=1}^{\mathrm{K}} P(\mathrm{word}_k|Y = c)^{x_k} && (2)\\
&= \pi_c \prod_{k=1}^{\mathrm{K}} \theta_{ck}^{x_k} && (3)
\end{aligned}
$$

where x_k is the number of occurrences of the kth word, π_c is the prior probability of class c (which allows multiple classes!), and θ_{ck} is the weight of the kth word for the cth class.

Learning multi-class naive Bayes

Training data

$$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{\mathsf{N}} \rightarrow \mathcal{D} = \{(\{x_{nk}\}_{k=1}^{\mathsf{K}}, y_n)\}_{n=1}^{\mathsf{N}}$$

Our goal

Learn $\pi_c, c = 1, 2, \cdots, \mathsf{C}$, and $\theta_{ck}, \forall c \in [\mathsf{C}], k \in [\mathsf{K}]$ under the constraints:

$$\sum_c \pi_c = 1$$

and

$$\sum_k \theta_{ck} = \sum_k P(\text{word}_k | Y = c) = 1$$

as well as π_c, $\theta_{ck} \geq 0$.

Our hammer: Maximum likelihood estimation

- Find the log-likelihood of the training data

$$
\begin{aligned}
\mathcal{L} = \log P(\mathcal{D}) &= \log \prod_{n=1}^{\mathsf{N}} \pi_{y_n} P(\mathbf{x}_n | y_n) \\
&= \log \prod_{n=1}^{\mathsf{N}} \left(\pi_{y_n} \prod_k \theta_{y_n k}^{x_{nk}} \right) \\
&= \sum_n \left(\log \pi_{y_n} + \sum_k x_{nk} \log \theta_{y_n k} \right) \\
&= \sum_n \log \pi_{y_n} + \sum_{n,k} x_{nk} \log \theta_{y_n k}
\end{aligned}
$$

- Optimize it!

$$
(\pi_c^*, \theta_{ck}^*) = \arg\max \sum_n \log \pi_{y_n} + \sum_{n,k} x_{nk} \log \theta_{y_n k}
$$

Our hammer: Maximum likelihood estimation

Optimization Problem

$$(\pi_c^*, \theta_{ck}^*) = \arg\max \left(\sum_n \log \pi_{y_n} + \sum_{n,k} x_{nk} \log \theta_{y_n k} \right)$$

Solution

$$\theta_{ck}^* = \frac{\#\text{of times word } k \text{ shows up in data points labeled as } c}{\#\text{total trials for data points labeled as } c}$$

$$\pi_c^* = \frac{\#\text{of data points labeled as c}}{\text{N}}$$

Outline

1. Review of Logistic Regression

2. Non-linear Decision Boundaries

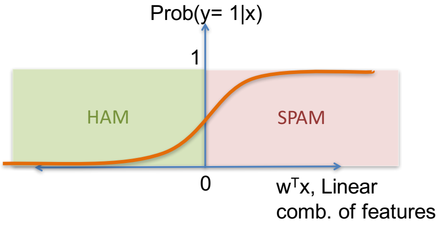
3. Multi-class Classification

   Multi-class Naive Bayes

   Multi-class Logistic Regression

Logistic regression for predicting multiple classes?

- The linear decision boundary that we optimized was specific to binary classification.
  - If $\sigma(\mathbf{w}^\top \mathbf{x}) \geq 0.5$ declare $y = 1$ (spam)
  - If $\sigma(\mathbf{w}^\top \mathbf{x}) < 0.5$ declare $y = 0$ (ham)
- How to extend it to multi-class classification?

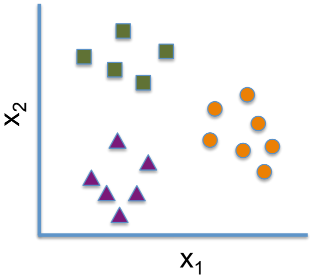


$y = 1$ for spam, $y = 0$ for ham

Idea: Express as multiple binary classification problems

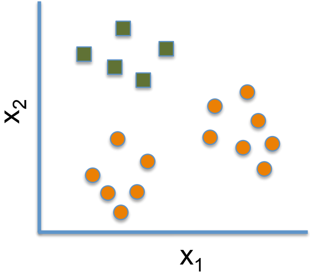
The One-versus-Rest or One-versus-All approach

- For each class c, change the problem into binary classification
  1. Relabel training data with label c, into POSITIVE (or '1').
  2. Relabel all the rest data into NEGATIVE (or '0').
- Repeat this multiple times: Train C binary classifiers, using logistic regression to differentiate the two classes each time.



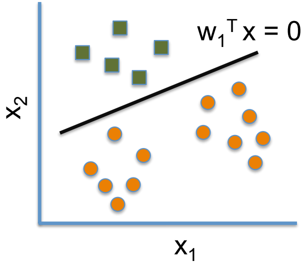
The One-versus-Rest or One-versus-All approach

- For each class c, change the problem into binary classification
  1. Relabel training data with label c, into POSITIVE (or '1')
  2. Relabel all the rest data into NEGATIVE (or '0')
- Repeat this multiple times: Train C binary classifiers, using logistic regression to differentiate the two classes each time.



The One-versus-Rest or One-versus-All approach

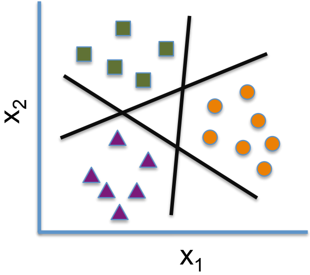
- For each class c, change the problem into binary classification
  1. Relabel training data with label c, into POSITIVE (or '1')
  2. Relabel all the rest data into NEGATIVE (or '0')
- Repeat this multiple times: Train C binary classifiers, using logistic regression to differentiate the two classes each time.



The One-versus-Rest or One-versus-All approach

How to combine these linear decision boundaries?

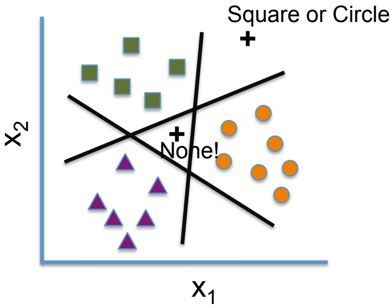
- There is ambiguity in some of the regions (the 4 triangular areas).



The One-versus-Rest or One-versus-All approach

How to combine these linear decision boundaries?

- There is ambiguity in some of the regions (the 4 triangular areas).
- How do we resolve this?

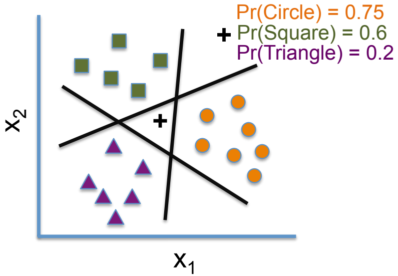


The One-versus-Rest or One-versus-All approach

How to combine these linear decision boundaries?

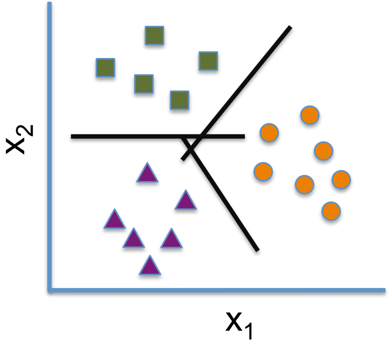
- Use the confidence estimates $\Pr(y = 1|\mathbf{x}) = \sigma(\mathbf{w}_1^\top \mathbf{x})$, $\ldots \Pr(y = C|\mathbf{x}) = \sigma(\mathbf{w}_C^\top \mathbf{x})$
- Declare class c^* that maximizes

$$c^* = \arg \max_{c=1,\ldots,C} \Pr(y = c|\mathbf{x}) = \sigma(\mathbf{w}_c^\top \mathbf{x})$$



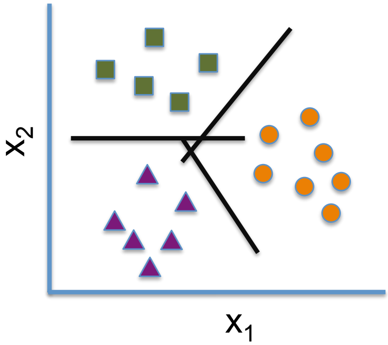
The One-versus-One approach

- For each **pair** of classes c and c', change the problem into binary classification.
  1. Relabel training data with label c, into POSITIVE (or '1')
  2. Relabel training data with label c' into NEGATIVE (or '0')
  3. **Disregard** all other data



The One-versus-One approach

- How many binary classifiers for C classes? $C(C-1)/2$
- How to combine their outputs?
- Given $\boldsymbol{x}$, count the $C(C-1)/2$ votes from outputs of all binary classifiers and declare the winner as the predicted class.
- Use confidence scores to resolve ties.



Contrast these approaches

Number of binary classifiers to be trained

- One-versus-All: C classifiers.
- One-versus-One: $C(C-1)/2$ classifiers – bad if C is large

Effect of relabeling and splitting training data

- One-versus-All: imbalance in the number of positive and negative samples can cause bias in each trained classifier.
- One-versus-One: each classifier trained on a small subset of data (only data in two classes), which can result in high variance.

Any other ideas?

- Hierarchical classification – we will see this in decision trees
- Multinomial logistic regression – directly output probabilities of y being in each of the C classes.

Multinomial logistic regression

Intuition:

from the decision rule of our naive Bayes classifier

$$\begin{aligned} y^* &= \arg\max_c P(y = c|\boldsymbol{x}) = \arg\max_c \log p(\boldsymbol{x}|y = c)p(y = c) \\ &= \arg\max_c \log \pi_c + \sum_k x_k \log \theta_{ck} = \arg\max_c \boldsymbol{w}_c^\top \boldsymbol{x} \end{aligned}$$

Essentially, we are comparing

$$\boldsymbol{w}_1^\top \boldsymbol{x}, \boldsymbol{w}_2^\top \boldsymbol{x}, \cdots, \boldsymbol{w}_{\mathsf{C}}^\top \boldsymbol{x}$$

with **one** for each category.

First try

So, can we define the following conditional model?

$$P(y = c|\boldsymbol{x}) = \sigma[\boldsymbol{w}_c^\top \boldsymbol{x}].$$

This would **not** work because:

$$\sum_c P(y = c|\boldsymbol{x}) = \sum_c \sigma[\boldsymbol{w}_c^\top \boldsymbol{x}] \neq 1,$$

so each summand can be any number (independently) between 0 and 1.

But we are close!

Learn the C linear models jointly to ensure this property holds!

Multinomial logistic regression

- **Model:** For each class c, we have a parameter vector $\boldsymbol{w}_c$ and model the posterior probability as:

$$P(c|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_c^\top \boldsymbol{x}}}{\sum_{c'} e^{\boldsymbol{w}_{c'}^\top \boldsymbol{x}}} \qquad \leftarrow \qquad \textit{This is called the softmax function.}$$

- **Decision boundary:** Assign $\boldsymbol{x}$ with the label that is the maximum of posterior:

$$\arg\max_c P(c|\boldsymbol{x}) \rightarrow \arg\max_c \boldsymbol{w}_c^\top \boldsymbol{x}.$$

How does the softmax function behave?

Suppose we have

$$\boldsymbol{w}_1^\top \boldsymbol{x} = 100, \quad \boldsymbol{w}_2^\top \boldsymbol{x} = 50, \quad \boldsymbol{w}_3^\top \boldsymbol{x} = -20.$$

We would pick the winning class label 1.

Softmax translates these scores into well-formed conditional probabilities

$$P(y = 1|\boldsymbol{x}) = \frac{e^{100}}{e^{100} + e^{50} + e^{-20}} < 1$$

- Preserves relative ordering of scores.
- Maps scores to values between 0 and 1 that also sum to 1.

Sanity check

Multinomial model reduces to binary logistic regression when $C = 2$.

$$P(1|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_1^\top \boldsymbol{x}}}{e^{\boldsymbol{w}_1^\top \boldsymbol{x}} + e^{\boldsymbol{w}_2^\top \boldsymbol{x}}} = \frac{1}{1 + e^{-(\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \boldsymbol{x}}}$$
$$= \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}}}$$

when we define $\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2$. Multinomial logistic regression thus generalizes the (binary) logistic regression to deal with multiple classes.

Parameter estimation for multinomial logistic regression

Discriminative approach: Maximize conditional likelihood

$$\log P(\mathcal{D}) = \sum_n \log P(y_n|\boldsymbol{x}_n)$$

We will change y_n to $\boldsymbol{y}_n = [y_{n1} \ y_{n2} \ \cdots \ y_{nC}]^\top$, a C-dimensional vector using 1-of-C encoding.

$$y_{nc} = \begin{cases} 1 & \text{if } y_n = c \\ 0 & \text{otherwise} \end{cases}$$

Ex: if $y_n = 2$, then, $\boldsymbol{y}_n = [0 \ 1 \ 0 \ 0 \ \cdots \ 0]^\top$.

$$\Rightarrow \sum_n \log P(y_n|\boldsymbol{x}_n) = \sum_n \log \prod_{c=1}^{C} P(c|\boldsymbol{x}_n)^{y_{nc}} = \sum_n \sum_c y_{nc} \log P(c|\boldsymbol{x}_n)$$

Cross-entropy error function

Definition: negative log-likelihood

$$\mathcal{E}(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C) = -\sum_n \sum_c y_{nc} \log P(c|\boldsymbol{x}_n)$$
$$= -\sum_n \sum_c y_{nc} \log \left(\frac{e^{\boldsymbol{w}_c^\top \boldsymbol{x}_n}}{\sum_{c'} e^{\boldsymbol{w}_{c'}^\top \boldsymbol{x}_n}} \right)$$

Properties of cross-entropy

- Convex in the **w** vectors, therefore unique global optimum
- Optimization requires numerical procedures, analogous to those used for binary logistic regression.

Finding the gradient

$$\mathcal{E}(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C) = -\sum_n \sum_c y_{nc} \log P(c|\boldsymbol{x}_n)$$
$$= -\sum_n \sum_c y_{nc} \log \left(\frac{e^{\boldsymbol{w}_c^\top \boldsymbol{x}_n}}{\sum_{c'} e^{\boldsymbol{w}_{c'}^\top \boldsymbol{x}_n}} \right)$$

- Need to find the gradient w.r.t. $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C$ and update

$$\boldsymbol{w}_c \leftarrow \boldsymbol{w}_c - \eta \frac{\partial \mathcal{E}}{\partial \boldsymbol{w}_c}, \qquad c = 1, \ldots, C$$

Can you find the gradient? (Hint: what is the gradient of the softmax function?)

Summary

You should know

- Differences between Naive Bayes and Logistic Regression.
- How to solve for the model parameters using gradient descent.
- How to generalize logistic regression to handle nonlinear decision boundaries.
- How to handle multiclass classification: one-versus-all, one-versus-one, multinomial regression.