

Naive Bayes and Logistic Regression

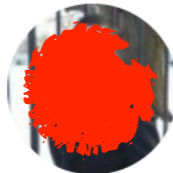
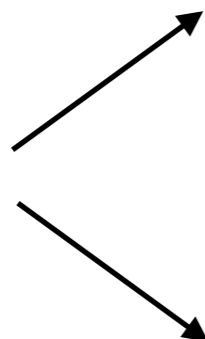
Nishant Mehta

Lecture 13

Text classification

Model problem: Classifying restaurant reviews as positive or negative


Training examples



[REDACTED]
Boston, United States
👤 82 🌟 24

★★★★★ 8/10/2019

Very reasonable priced single origin Bolivian. I had the pour over, which had distinct tasting notes of almond and plum. The staff were polite and friendly.



[REDACTED]
Pioneer, United States
👤 5 🌟 74 📷 191

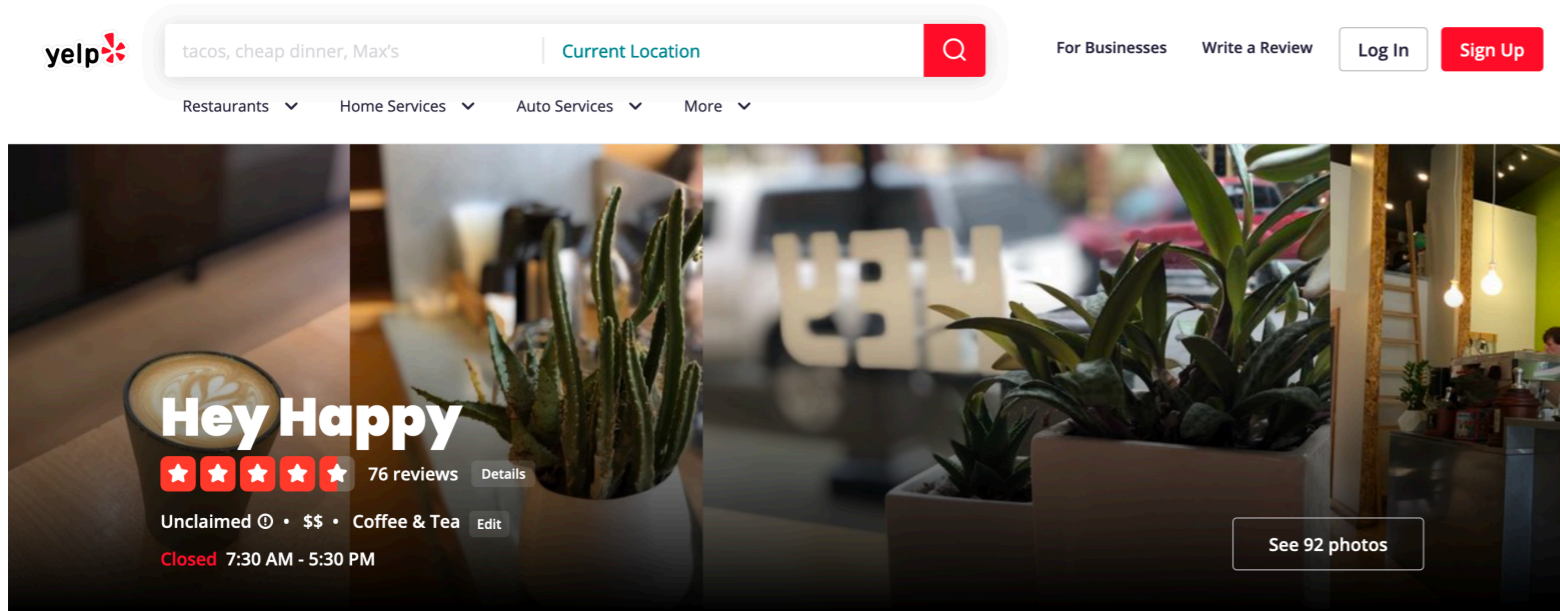
★★★★★ 11/23/2018 · 🔄 Updated review

📷 1 photo

Nitro cold brew?? Oh yes. What an amazing glass of coffee. Watching it being poured was like watching the pouring of a fine pint of Guinness . It was smooth with a hint of effervescence. Truly memorable.

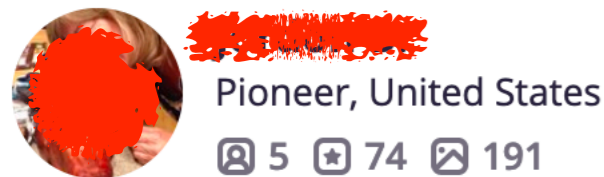
How to represent data?

Bag of Words Representation



Boolean value indicating whether word occurs in document

Subset of words in the vocabulary



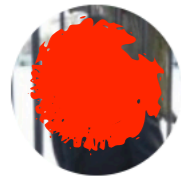
★★★★★ 11/23/2018 · Updated review

1 photo

Nitro cold brew?? Oh yes. What an amazing glass of coffee. Watching it being poured was like watching the pouring of a fine pint of Guinness. It was smooth with a hint of effervescence. Truly memorable.

coffee	1
espresso	0
amazing	1
drink	0
guinness	1
nitro	1
latte	0
⋮	⋮

Positive examples



[Redacted Name]
Boston, United States
82 24

★★★★★ 8/10/2019

Very reasonable priced single origin Bolivian. I had the pour over, which had distinct tasting notes of almond and plum. The staff were polite and friendly.



[Redacted Name]
Pioneer, United States
5 74 191

★★★★★ 11/23/2018 · Updated review

1 photo

Nitro cold brew?? Oh yes. What an amazing glass of coffee. Watching it being poured was like watching the pouring of a fine pint of Guinness . It was smooth with a hint of effervescence. Truly memorable.



[Redacted Name]
Lahaina, United States
0 61 8

★★★★★ 7/3/2019

Barista champion shop! Amazingggggg ristretto and cappuccino for you with refined coffee palettes. What a find. Fantastic location. We rode our bikes around town and did our own coffee tour. What a treat.

Positive Words

reasonable

distinct

polite

friendly

amazing

fine

effervescence

memorable

amazingggggg

fantastic

treat

Negative examples



[Redacted]

Vancouver, BC

6 115 24



12/23/2017

1 photo

So I was craving some java and sweet treats. Came here on an afternoon because it seems recommended on yelp..

Ambience and decor is nice. Seating is limited and staff was friendly. I ordered a chai latte with an espresso on the side and a power cookie.

The power cookie was great. The rest horrid.

One of the best ways to judge a coffee place is by just having an espresso. It was rude. Like sipping bitter vinegar. Absolutely nasty..

The chai was slightly better but compared to the Whole foods Chai's it did not measure up. Bland, too much milk, and while presented nice it was all show.

Would have to see if there is anything else worth it otherwise not a good first impression

Negative words

limited

horrid

rude

bitter

nasty

bland

Predicting the most likely class

Suppose we model probability of label Y given input feature vector X

$$P(Y = y \mid X = x, \theta)$$

Natural rule: Predict most likely label according to our model

$$\hat{y} = \arg \max_{y \in \mathcal{Y}} P(Y = y \mid X = x, \theta)$$

How to model $P(Y \mid X)$? We use Bayes rule:

$$\begin{aligned} P(Y = y \mid X = x, \theta) &= \frac{P(X = x \mid Y = y, \theta)P(Y = y \mid \theta)}{P(X = x \mid \theta)} \\ &= \frac{P(X = x \mid Y = y, \theta)P(Y = y \mid \theta)}{\sum_{y' \in \mathcal{Y}} P(X = x \mid Y = y', \theta)P(Y = y' \mid \theta)} \end{aligned}$$

Generative model

A *generative model* is based on modeling the full joint distribution $P(X, Y)$

$$P(Y = y \mid X = x, \theta) = \frac{P(X = x \mid Y = y, \theta)P(Y = y \mid \theta)}{\sum_{y' \in \mathcal{Y}} P(X = x \mid Y = y', \theta)P(Y = y' \mid \theta)}$$

Two steps:

Model the prior probability of any example having class y

Model the probability distribution of examples from class y

The power of generative models

With a model of $P(X | Y)$, we can generate new examples



Brock et al. (DeepMind)

A closer look at $P(X | Y)$

Suppose we have d input features, each taking J possible values.

So, for any y , $P(X | Y = y, \theta)$ is a categorical distribution over outcomes (one outcome per feature vector!)

θ is a probability vector - how many parameters?

Recall the restaurant review classification problem...

Bag of words encoding:

$d = 1000$ (reasonable vocabulary size)

$J = 2$ (boolean features)

A closer look at $P(X | Y)$


Suppose we have d input features, each taking J possible values.

So, for any y , $P(X | Y = y, \theta)$ is a categorical distribution over J^d outcomes (one outcome per feature vector!)

θ is a probability vector - how many parameters? $J^d - 1$ for each label y

J^d outcomes

$$(\hat{\theta}_{MLE})_j = \frac{n_j}{n}$$

$n_j = 0$ for all but at most
 $J^d - n$ outcomes j

$K(J^d - 1)$ in total
number of labels

Recall the restaurant review classification problem...

Bag of words encoding:

$d = 1000$ (reasonable vocabulary size)

$J = 2$ (boolean features)

Naive Bayes assumption

Naive Bayes assumption: features are independent given the label

$$P((X_1, \dots, X_d) = (x_1, \dots, x_d) \mid Y = y)$$

Naive Bayes assumption

Naive Bayes assumption: features are independent given the label

$$\begin{aligned} &P((X_1, \dots, X_d) = (x_1, \dots, x_d) \mid Y = y) \\ &= P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \cdot \dots \cdot P(X_d = x_d \mid Y = y) \\ &= \prod_{j=1}^d P(X_j = x_j \mid Y = y) \end{aligned}$$

Naive Bayes assumption

Naive Bayes assumption: features are independent given the label

$$\begin{aligned} &P((X_1, \dots, X_d) = (x_1, \dots, x_d) \mid Y = y) \\ &= P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \cdot \dots \cdot P(X_d = x_d \mid Y = y) \\ &= \prod_{j=1}^d P(X_j = x_j \mid Y = y) \end{aligned}$$

How many parameters are needed to model $P(X_1 = x_1 \mid Y = y)$?

How many parameters are needed in total?

Naive Bayes assumption

Naive Bayes assumption: features are independent given the label

$$\begin{aligned} &P((X_1, \dots, X_d) = (x_1, \dots, x_d) \mid Y = y) \\ &= P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \cdot \dots \cdot P(X_d = x_d \mid Y = y) \\ &= \prod_{j=1}^d P(X_j = x_j \mid Y = y) \end{aligned}$$

How many parameters are needed to model $P(X_1 = x_1 \mid Y = y)$? $J - 1$

How many parameters are needed in total?

Naive Bayes assumption

Naive Bayes assumption: features are independent given the label

$$\begin{aligned} P((X_1, \dots, X_d) = (x_1, \dots, x_d) \mid Y = y) \\ &= P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \cdot \dots \cdot P(X_d = x_d \mid Y = y) \\ &= \prod_{j=1}^d P(X_j = x_j \mid Y = y) \end{aligned}$$

How many parameters are needed to model $P(X_1 = x_1 \mid Y = y)$? $J - 1$

How many parameters are needed in total? $(J - 1)dK$

 number of labels

Naive Bayes classifier

Naive Bayes classifier predicts

$$\begin{aligned}\hat{y} &= \arg \max_{k \in \{1, 2, \dots, K\}} P(Y = k \mid (X_1, \dots, X_d) = (x_1, \dots, x_d), \theta) \\ &= \arg \max_{k \in \{1, 2, \dots, K\}} \frac{P(Y = k \mid \theta) \prod_{i=1}^d P(X_i = x_i \mid Y = k, \theta)}{\sum_{j=1}^K P(Y = j \mid \theta) \prod_{i=1}^d P(X_i = x_i \mid Y = j, \theta)}\end{aligned}$$

What is the form of the log likelihood?

We have Data $D = ((X_1, Y_1), \dots, (X_n, Y_n))$

The log likelihood is

$$\begin{aligned}& \sum_{i=1}^n \log P(X_{i,1} = x_{i,1}, \dots, X_{i,d} = x_{i,d}, Y_i = y_i \mid \theta) \\ &= \sum_{i=1}^n \left(\log P(Y_i = y_i \mid \theta) + \sum_{j=1}^d \log P(X_{i,j} = x_{i,j} \mid Y_i = y_i, \theta) \right)\end{aligned}$$

How to estimate parameter?

Idea 1: Use MLE

Recall: Restaurant review classification 'f' for "fantastic"

$$P(X_{\text{fantastic}} = 1 \mid Y = +1, \theta_{\text{pos},f}) = (\theta_{\text{pos},f})^{n_{\text{pos},f}} (1 - \theta_{\text{pos},f})^{n_{\text{pos},\bar{f}}}$$

Maximize WRT $\theta_{\text{pos},f}$ just like MLE with Bernoulli distribution

$$\hat{\theta}_{\text{pos},f} = \frac{n_{\text{pos},f}}{n_{\text{pos}}} = \frac{n_{\text{pos},f}}{n_{\text{pos},f} + n_{\text{pos},\bar{f}}}$$

... But what if that review with `amazingggggg` only occurs in test set?

Then from MLE, $P(X_{\text{amazingggggg}} = 1 \mid Y = +1, \theta_{\text{pos},a}) = 0$

and also likewise $P(X_{\text{amazingggggg}} = 1 \mid Y = -1, \theta_{\text{neg},a}) = 0$.

So, $P(Y = 1 \mid X, \hat{\theta}_{\text{MLE}}) = P(Y = 0 \mid X, \hat{\theta}_{\text{MLE}}) = 0$

Idea 2: Use add-one smoothing to fit individual model parameters

Discriminative models

Is Naive Bayes classifier good enough?

Naive Bayes assumption greatly reduces number of parameters

High bias (when Naive Bayes assumption is violated)

(But also an upside: low variance; more on this later...)



What if we try to estimate $P(Y | X)$ directly, using a linear model?

Discriminative model

Directly focus on discriminating label Y given input X

Solve the simplest task that we wish to solve (can have lower bias as we don't even try to estimate $P(X | Y)$)

Logistic regression

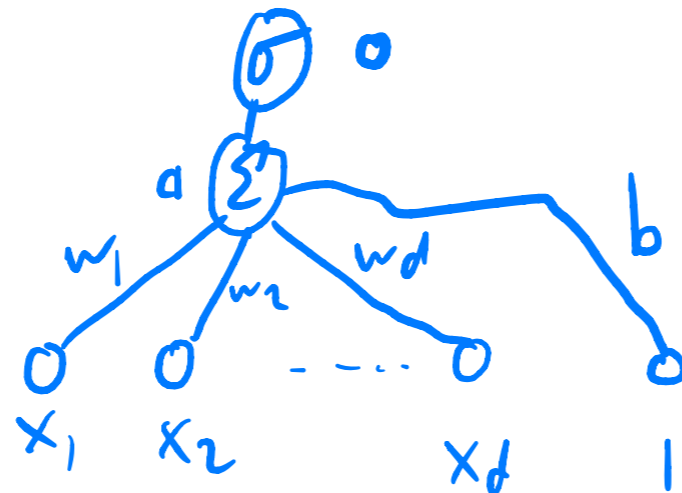
Consider binary classification with labels $y \in \{0, 1\}$ and feature vectors $x \in \mathbb{R}^d$

Parameters: weights vector $w \in \mathbb{R}^d$ and bias term $b \in \mathbb{R}$

Logistic regression predictor takes form:

$$f_{w,b}(x) = \sigma(\langle w, x \rangle + b) = \frac{1}{1 + e^{-\langle w, x \rangle + b}}$$

Neural network diagram representation:



Conditional probabilities

$$P(Y = 1 \mid X = x, w) = \sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)} = \frac{\exp(\langle w, x \rangle)}{1 + \exp(\langle w, x \rangle)}$$

$$P(Y = 0 \mid X = x, w) = 1 - \sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(\langle w, x \rangle)}$$

Unified form: $P(Y = y \mid X = x, w) = \underbrace{\sigma(\langle w, x \rangle)^y}_{\text{"}\theta^y\text{"}} \underbrace{(1 - \sigma(\langle w, x \rangle))^{1-y}}_{\text{"}(1-\theta)^{1-y}"}$

analogy with
MLE for
Bernoulli distribution
with parameter θ

θ depends on x

Logistic regression - Intuition

$$P(Y = 1 \mid X = x, w) = \sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)} = \frac{\exp(\langle w, x \rangle)}{1 + \exp(\langle w, x \rangle)}$$

$$P(Y = 0 \mid X = x, w) = 1 - \sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(\langle w, x \rangle)}$$

Where does the logistic regression functional form come from?

How to fit model? Use maximum conditional likelihood estimation

$$\text{Recall } P(Y = y \mid X = x, w) = \sigma(\langle w, x \rangle)^y (1 - \sigma(\langle w, x \rangle))^{1-y}$$

$$\max_w \log P(y_1, y_2, \dots, y_n \mid x_1, x_2, \dots, x_n, w)$$

$$= \max_w \log \prod_{j=1}^n P(y_j \mid x_j, w)$$

$$= \max_w \sum_{j=1}^n \log P(y_j \mid x_j, w)$$

$$= \max_w \sum_{j=1}^n \log \left[\sigma(\langle w, x_j \rangle)^{y_j} (1 - \sigma(\langle w, x_j \rangle))^{1-y_j} \right]$$

$$= \max_w \sum_{j=1}^n \left[y_j \log \sigma(\langle w, x_j \rangle) + (1-y_j) \log (1 - \sigma(\langle w, x_j \rangle)) \right]$$

$f_w(x_j) \leftarrow$ probability predictor

Equivalent: \min_w

$$\sum_{j=1}^n \left[-y_j \log f_w(x_j) - (1-y_j) \log (1 - f_w(x_j)) \right]$$

$\mathcal{L}(y_j, f_w(x_j)) \leftarrow$ cross-entropy loss

Extension to multi-class classification

What if we have K classes?

For each class label j , maintain a separate parameter vector $w_j \in \mathbb{R}^d$

Parameter can be viewed as a matrix $W = (w_1 \ w_2 \ \cdots \ w_K) \in \mathbb{R}^{d \times K}$

Now, the conditional probability of class j given $X = x$ is modeled as

$$P(Y = j \mid X = x, W) = \frac{\exp(\langle w_j, x \rangle)}{\sum_{c=1}^K \exp(\langle w_c, x \rangle)}$$


Extension to multi-class classification

What if we have K classes?

For each class label j , maintain a separate parameter vector $w_j \in \mathbb{R}^d$

Parameter can be viewed as a matrix $W = (w_1 \ w_2 \ \cdots \ w_K) \in \mathbb{R}^{d \times K}$

Now, the conditional probability of class j given $X = x$ is modeled as

$$P(Y = j \mid X = x, W) = \frac{\exp(\langle w_j, x \rangle)}{\sum_{c=1}^K \exp(\langle w_c, x \rangle)}$$


Softmax: given z_1, z_2, \dots, z_K , transform z_j as $z_j \mapsto \frac{\exp(z_j)}{\sum_{c=1}^K \exp(z_c)}$

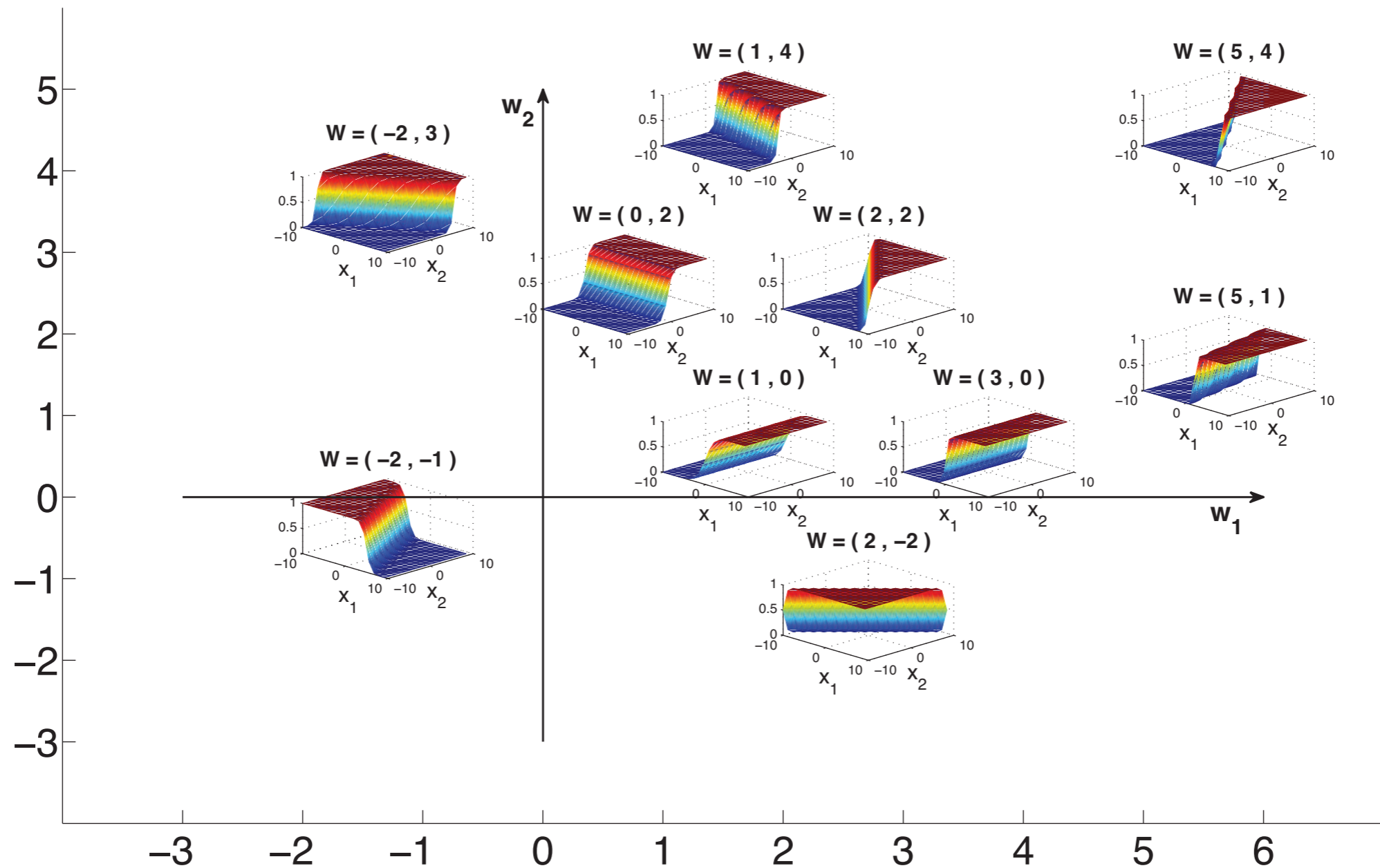


Figure 8.1 Plots of $\text{sigm}(w_1x_1 + w_2x_2)$. Here $\mathbf{w} = (w_1, w_2)$ defines the normal to the decision boundary. Points to the right of this have $\text{sigm}(\mathbf{w}^T \mathbf{x}) > 0.5$, and points to the left have $\text{sigm}(\mathbf{w}^T \mathbf{x}) < 0.5$. Based on Figure 39.3 of (MacKay 2003). Figure generated by `sigmoidplot2D`.

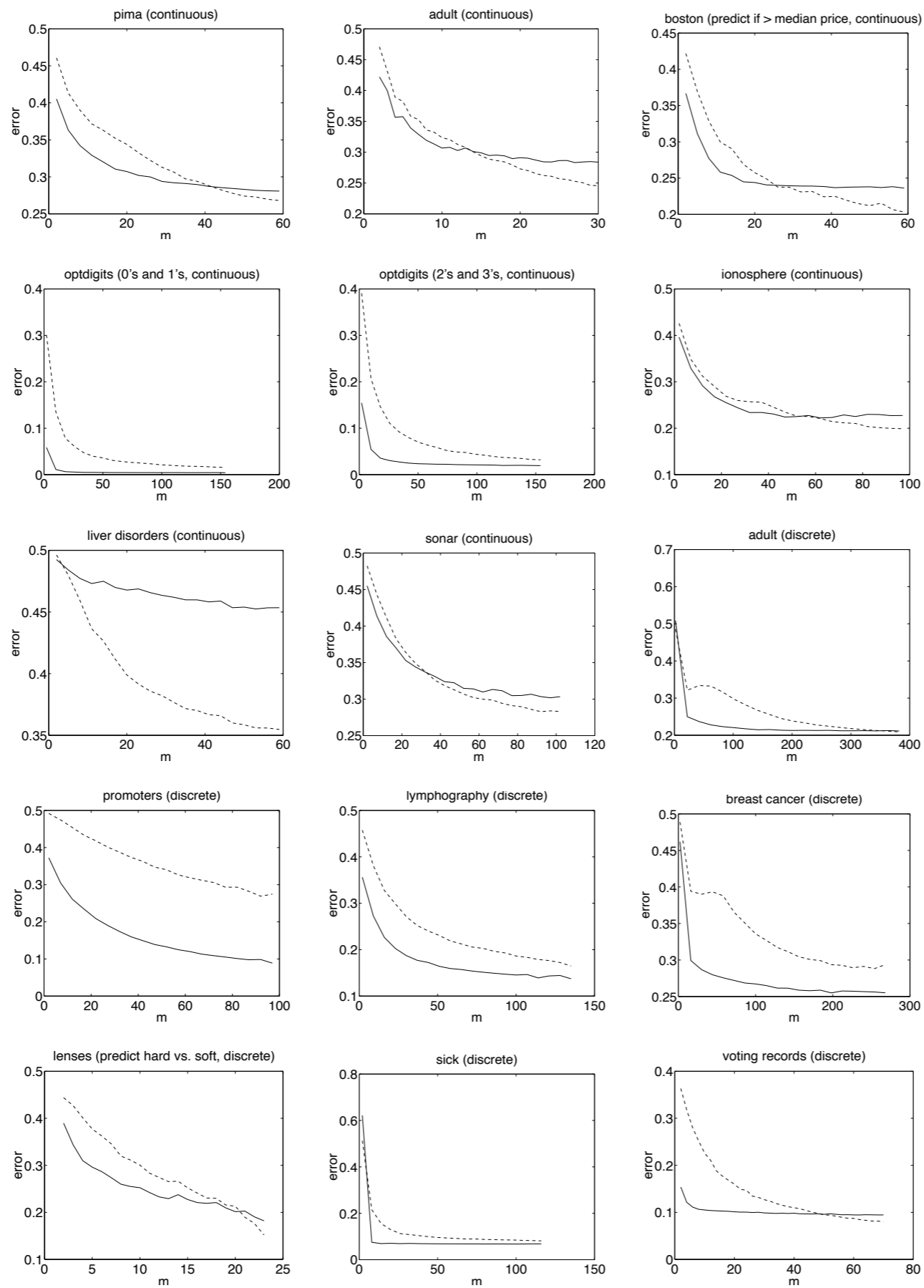


Figure 1: Results of 15 experiments on datasets from the UCI Machine Learning repository. Plots are of generalization error vs. m (averaged over 1000 random train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

(Ng and Jordan, 2001)
 “On Discriminative vs. Generative classifiers:
 A comparison of logistic regression and
 naive Bayes”