LECTURE 2

Supervised text classification

SUPERVISED TEXT CLASSIFICATION

- Objective
  - Given a set of annotated texts, *i.e.* examples
    - $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \quad \forall (x_i, y_i)_{1 \leq i \leq n} : x_i \in \mathcal{X}, y_i \in \mathcal{Y}$
  - Find a function (the decision rule) that predicts the class of a document
    - $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Lecture plan
  - Two parametric models
    - Multinomial naïve Bayse classifier
    - Logistic regression

MULTINOMIAL NAÏVE BAYES CLASSIFIER

MODEL

- Data: a set of annotated documents
  - $D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \forall (x_i, y_i)_{1 \leq i \leq n} : x_i \in \mathbb{N}^d, y_i \in \{1, \ldots, q\}$
- Joint distribution
  - $P(X = x, Y = y) = P(Y = y)P(X = x \mid Y = y)$
- Naïve independence assumption
  - $P(X = x \mid Y = y) = \prod_{j=1}^{d} P(X_j = x_j \mid Y = y)$
  - $P(X = x, Y = y) = P(Y = y) \prod_{j=1}^{d} P(X_j = x_j \mid Y = y)$

MODEL

- Decision rule
  - $\hat{y} = \underset{y \in \{1,2,\ldots,q\}}{\operatorname{argmax}} \left(P(Y = y \,|\, X = z) \right)$
  - By Bayes rule we have
    - $\hat{y} = \underset{y \in \{1,2,\ldots,q\}}{\operatorname{argmax}} \left(\frac{P(X = x \,|\, Y = y) P(Y = y)}{P(X = x)} \right)$
    - $\hat{y} = \underset{y \in \{1,2,\ldots,q\}}{\operatorname{argmax}} \left(P(X = x \,|\, Y = y) P(Y = y) \right)$
- Introducing parameters we get
  - $\hat{y} = \underset{y \in \{1,2,\ldots,q\}}{\operatorname{argmax}} \left(g(y) \prod_{j=1}^{d} g_j(1 \,|\, y)^{z_j} \right)$

PARAMETER ESTIMATION

- Maximum likelihood estimation (MLE)
  - $L(D; g, g_j) = \prod_{i=1}^{n} P(X = x_i, Y = y_i)$
  - $\ell(D; g, g_j) = \log\left(\prod_{i=1}^{n} P(X = x_i, Y = y_i) \right)$

$$= \log\left(\prod_{i=1}^{n} \left(g(y_i) \prod_{j=1}^{d} g_j(x_{ij} \mid y_i) \right) \right)$$

$$= \sum_{i=1}^{n} \log\left(g(y_i) \prod_{j=1}^{d} g_j(x_{ij} \mid y_i) \right)$$

$$= \sum_{i=1}^{n} \left[\log\left(g(y_i) \right) + \sum_{j=1}^{d} \log g_j(x_{ij} \mid y_i) \right]$$

PARAMETER ESTIMATION

- Maximum likelihood estimation (MLE)
  - $g(y)$ is the relative frequency of documents of class y in D
    - $g(y) = \frac{\sum_{i=1}^{n} \delta_{y_i=y}}{n}$
  - $g_j(1 \mid y)$ is the relative frequency of word j in documents belonging to class y
    - $g_j(1 \mid y) = \frac{\sum_{i=1}^{n} \delta_{y_i=y} x_{ij}}{\sum_{i=1}^{n} \sum_{k=1}^{d} \delta_{y_i=y} x_{ik}}$

LAPLACE SMOOTHING

- MLE leads to zero estimates for some words, which is problematic given the decision rule
  - $\hat{y} = \underset{y \in \{1,2,\ldots,q\}}{\text{argmax}} \left(g(y) \prod_{j=1}^{d} g_j(1 \mid y)^{z_j} \right)$
- A usual workaround is applying Laplace smoothing
  - Simply increment the frequency of each word in each doc
  - $g_j(1 \mid y) = \dfrac{\sum_{i=1}^{n} \left(\delta_{y_i = y} x_{ij} + 1 \right)}{\sum_{i=1}^{n} \sum_{k=1}^{d} \left(\delta_{y_i = y} x_{ik} + 1 \right)}$

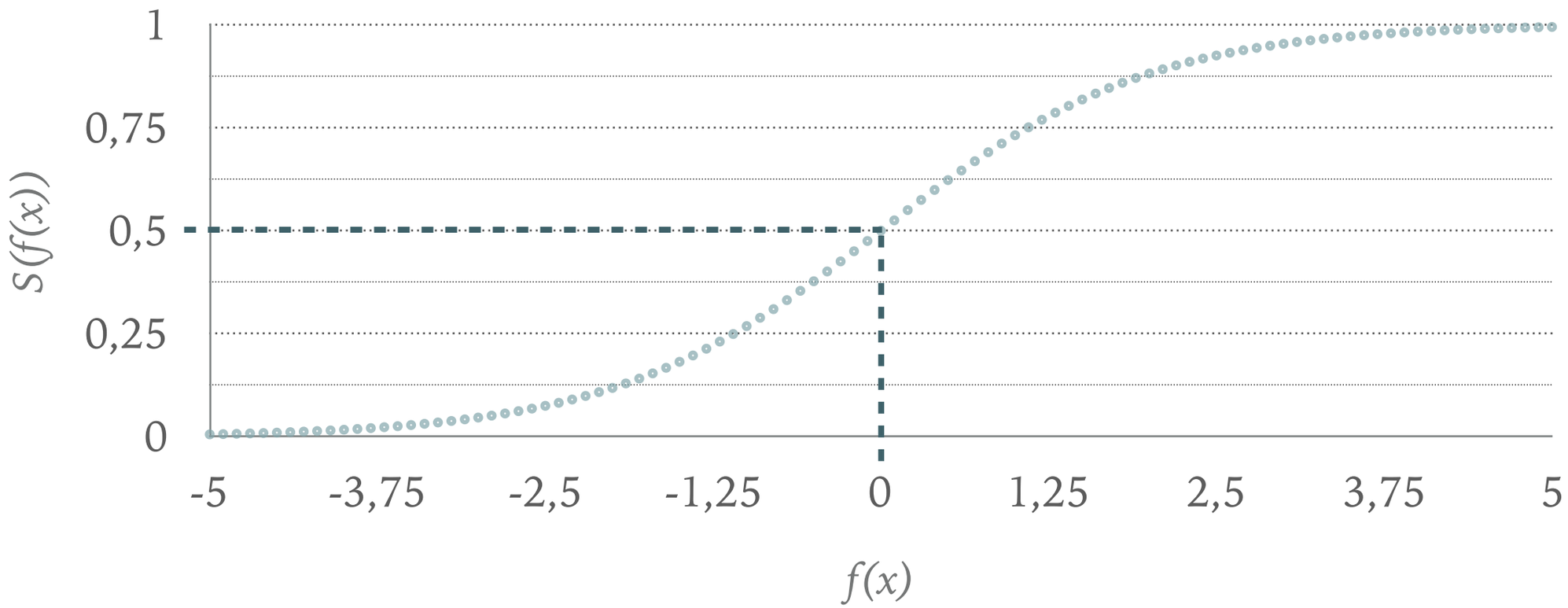
LOGISTIC REGRESSION

MODEL

- Data: a set of annotated documents
  - $D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \forall (x_i, y_i)_{1 \leq i \leq n} : x_i \in \mathbb{R}^d, y_i \in \{0,1\}$
- Conditional distribution
  - $P(Y = 1 \,|\, X = x) = \dfrac{1}{1 + e^{-f(x)}}$
    - We recognize the standard sigmoid function
      - $S : \mathbb{R} \to [0; 1] \qquad \alpha \mapsto \dfrac{1}{1 + e^{-\alpha}}$
    - Applied to a linear function of x, parameterized by w
      - $f : \mathbb{R}^d \to \mathbb{R}$
        $x \mapsto w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$

MODEL

- Decision rule
  - $\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} \Big(P(Y = y \,|\, X = z) \Big)$
- Linear classifier
  - Hyperplane equation $f(x) = 0$



PARAMETER ESTIMATION

- Maximum likelihood estimation
  - $L(D; w) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i)$

$$= \prod_{i=1}^{n} p(1 \mid x_i)^{y_i} (1 - p(1 \mid x_i))^{1-y_i}$$

  - $\ell(D; w) = \log\left(\prod_{i=1}^{n} p(1 \mid x_i)^{y_i} (1 - p(1 \mid x_i))^{1-y_i} \right)$

$$= \sum_{i=1}^{n} y_i \log\left(p(1 \mid x_i)\right) + (1 - y_i) \log\left((1 - p(1 \mid x_i)\right)$$

- Problem formulation
  - $w^* = \underset{w}{\operatorname{argmax}}(\ell(D; w)) = \underset{w}{\operatorname{argmax}}(L(D; w))$

PARAMETER ESTIMATION

- Rewrite the model and incorporate a constant $x_0 = 1$ so that
  - $$P(Y = 1 \mid X = x) = \frac{1}{1 + e^{-wx}} = S(wx)$$
- Partial derivative of the log-likelihood in w_j
  - $$\frac{\partial}{\partial w_j} \ell(D; w) = \frac{\partial}{\partial w_j} \sum_{i=1}^{n} y_i \log(p(1 \mid x_i)) + (1 - y_i)\log((1 - p(1 \mid x_i)))$$
  $$= \frac{\partial}{\partial w_j} \sum_{i=1}^{n} y_i \log(S(wx_i)) + (1 - y_i)\log((1 - S(wx_i)))$$
  $$= \sum_{i=1}^{n} y_i \frac{\partial}{\partial w_j} \log(S(wx_i)) + (1 - y_i)\frac{\partial}{\partial w_j} \log((1 - S(wx_i)))$$

PARAMETER ESTIMATION

- Partial derivative of the left-hand side of the log-likelihood
  - $$\frac{\partial}{\partial w_j} \log(S(wx_i)) = \frac{\partial}{\partial w_j} \log(\frac{1}{1+e^{-wx_i}})$$

$$= \frac{\partial}{\partial w_j} \log(1) - \log(1+e^{-wx_i})$$

$$= \frac{\partial}{\partial w_j} - \log(1+e^{-wx_i})$$

$$= -\frac{-x_{ij}e^{-wx_i}}{1+e^{-wx_i}}$$

$$= x_{ij}(1 - \frac{1}{1+e^{-wx_i}})$$

$$= x_{ij}(1 - S(wx_i))$$

PARAMETER ESTIMATION

- Partial derivative of the right-hand side of the log-likelihood
  - $$\frac{\partial}{\partial w_j} \log(1 - S(wx_i)) = \frac{\partial}{\partial w_j} \log(1 - \frac{1}{1 + e^{-wx_i}})$$
$$= \frac{\partial}{\partial w_j} \log(\frac{1 + e^{-wx_i}}{1 + e^{-wx_i}} - \frac{1}{1 + e^{-wx_i}})$$
$$= \frac{\partial}{\partial w_j} \log(\frac{e^{-wx_i}}{1 + e^{-wx_i}})$$
$$= \frac{\partial}{\partial w_j} \log(e^{-wx_i}) - \log(1 + e^{-wx_i})$$
$$= \frac{\partial}{\partial w_j} - wx_i - \log(1 + e^{-wx_i})$$
$$= - x_{ij} + x_{ij}(1 - S(wx_i))$$
$$= - x_{ij}S(wx_i)$$

PARAMETER ESTIMATION

- Assemble the two partial derivatives
  - $$\frac{\partial}{\partial w_j} \ell = \sum_{i=1}^{n} y_i x_{ij}(1 - S(wx_i)) - x_{ij}(1 - y_i)S(wx_i)$$

$$= \sum_{i=1}^{n} y_i x_{ij} - y_i x_{ij} S(wx_i) - x_{ij} S(wx_i) + y_i x_{ij} S(wx_i)$$

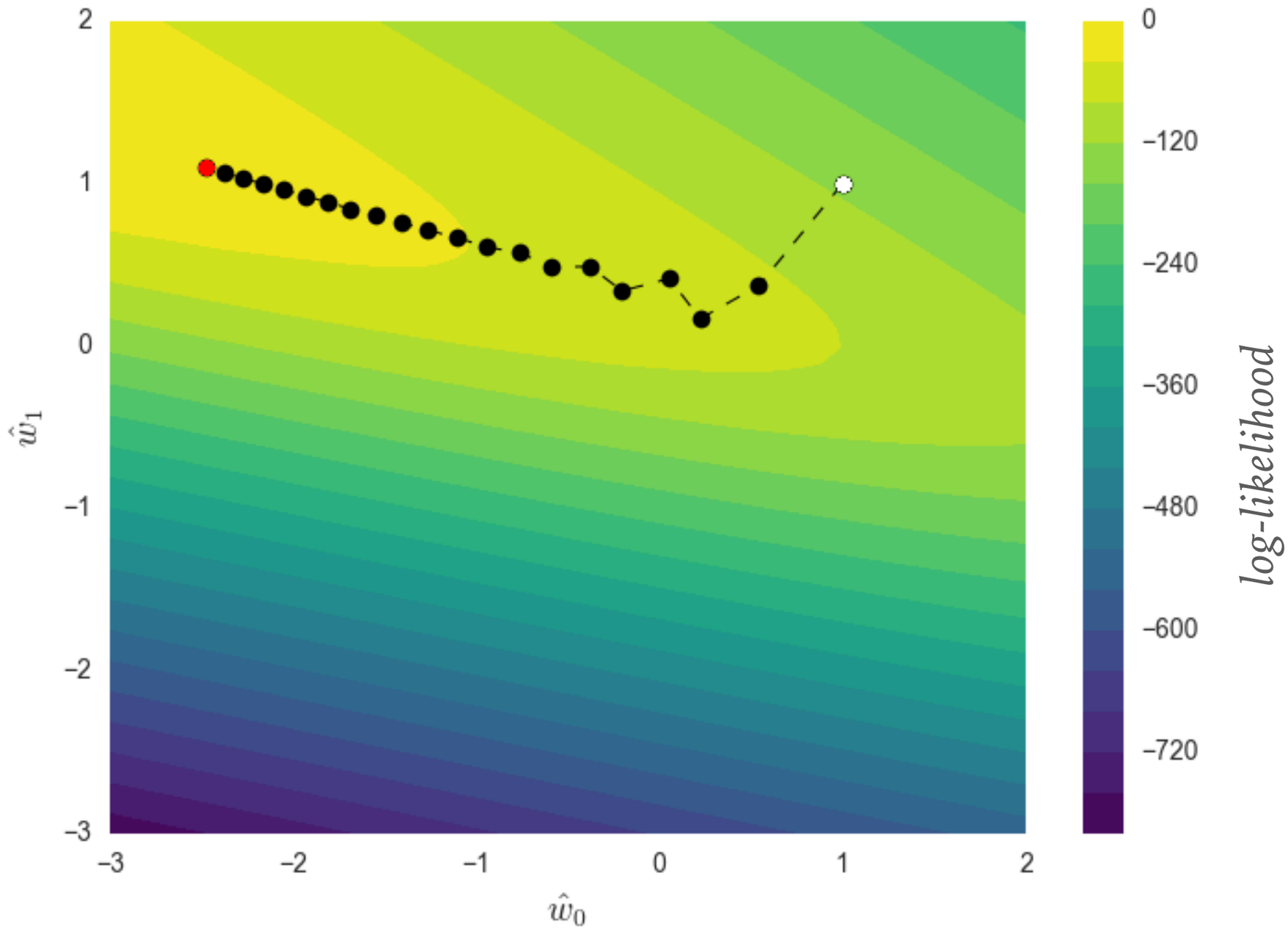
$$= \sum_{i=1}^{n} y_i x_{ij} - x_{ij} S(wx_i)$$

$$= \sum_{i=1}^{n} x_{ij}(y_i - S(wx_i))$$

- There's no analytic solution to $\frac{\partial}{\partial w_j} \ell(D; w) = 0$

APPROXIMATION VIA GRADIENT DESCENT

- Simple example of gradient descent



REGULARIZATION

- Add a penalization term, r, to the function to maximize
  - $w^* = \underset{w}{\text{argmax}}\left(\ell(D; w) - \lambda r(w)\right)$
- Common penalization terms
  - Ridge
    - $w^* = \underset{w}{\text{argmax}}\left(\ell(D; w) - \lambda \|w\|_2^2\right)$ with $\|w\|_2^2 = \sum_{j=0}^{d} w_j^2$
    - Penalize large coefficients
  - LASSO
    - $w^* = \underset{w}{\text{argmax}}\left(\ell(D; w) - \lambda \|w\|_1\right)$ with $\|w\|_1 = \sum_{j=0}^{d} |w_j|$
    - Favor sparse solutions

EVALUATION

CROSS-VALIDATION

- Split the data into two sets
  - Training set
  - Test set
- Methodology
  - Estimate the model parameters on the train set
  - Predict on the test set
  - Build the confusion matrix, *M*, *i.e.* a contingency table to compare predicted labels and true labels
- Repeat this *k* times with different training and test sets
  - *k*-fold cross-validation

CROSS-VALIDATION

- Common evaluation metrics
  - Overall accuracy
    - $A = \frac{\#\ of\ correctly\ classified\ documents}{total\ \#\ of\ documents}$
  - Precision w.r.t a given class
    - $P_c = \frac{\#\ of\ documents\ correctly\ classified\ in\ c}{\#\ of\ documents\ classified\ in\ c}$
  - Recall w.r.t a given class
    - $R_c = \frac{\#\ of\ documents\ correctly\ classified\ in\ c}{\#\ of\ documents\ belonging\ to\ class\ c}$

TAKE AWAY MESSAGE

- The multinomial naïve Bayes classifier is a good baseline, always try it first
  - Efficient parameter estimation
  - Sensitive to vocabulary filtering
  - Limited to multinomial data, *i.e.* raw word counts
- The classifier based on the logistic regression is more versatile
  - Compatible with any weighting scheme
  - Useful regularization
  - Higher computational complexity
- Both are linear classifiers