Multiclass linear prediction

CSE 250B

Multiclass classification

Of the classification methods we have studied so far, which seem inherently binary?

- Nearest neighbor?
- Generative models?
- Linear classifiers?

The main idea

Remember Gaussian generative models...

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$: specify a classifier by $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j|x) \;\propto\; e^{w_j \cdot x + b_j}$$

From binary to multiclass logistic regression

Binary logistic regression: for $\mathcal{X} = \mathbb{R}^d$, classifier given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr(y = 1|x) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$

Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$: specify a classifier by $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j|x) \propto e^{w_j \cdot x + b_j}$$

$\Pr(y=j|x) = C e^{\langle w_j, x \rangle + b_j}$

$\sum_j \Pr(y=j|x) = C \sum_j e^{\langle w_j, x \rangle + b_j} = 1.$

$C = \frac{1}{\sum_{j=1}^{k} e^{\langle w_j, x \rangle + b_j}}$

- What is the fully normalized form of the probability?

- Given a point x, which label to predict? $i = \underset{j \in \{1, \ldots, k\}}{\operatorname{argmax}} \Pr(y=j|x)$

Multiclass logistic regression

- **Label space**: $\mathcal{Y} = \{1, 2, \ldots, k\}$
- **Parametrized classifier**: $w_1, \ldots, w_k \in \mathbb{R}^d$, $b_1, \ldots, b_k \in \mathbb{R}$:

$$\Pr(y = j|x) = \frac{e^{w_j \cdot x + b_j}}{e^{w_1 \cdot x + b_1} + \cdots + e^{w_k \cdot x + b_k}}$$

- **Prediction**: given a point x, predict label $\arg\max_j (w_j \cdot x + b_j)$.

$$\arg\max_j e^{\langle w_j, x\rangle + b_j} = \arg\max_j \langle w_j, x\rangle + b_j$$

- **Learning**: Given: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$.
  Find: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k$ that maximize the likelihood

$$\max \prod_{i=1}^{n} \Pr(y^{(i)}|x^{(i)}) \equiv \min -\log P(y^{(i)}|x^{(i)})$$

  Taking negative log gives a convex minimization problem.

Multiclass Perceptron

Setting: $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \ldots, k\}$

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Multiclass Perceptron

Setting: $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \ldots, k\}$

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

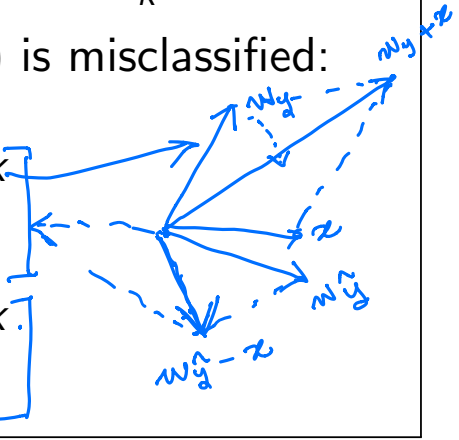
Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Learning. Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

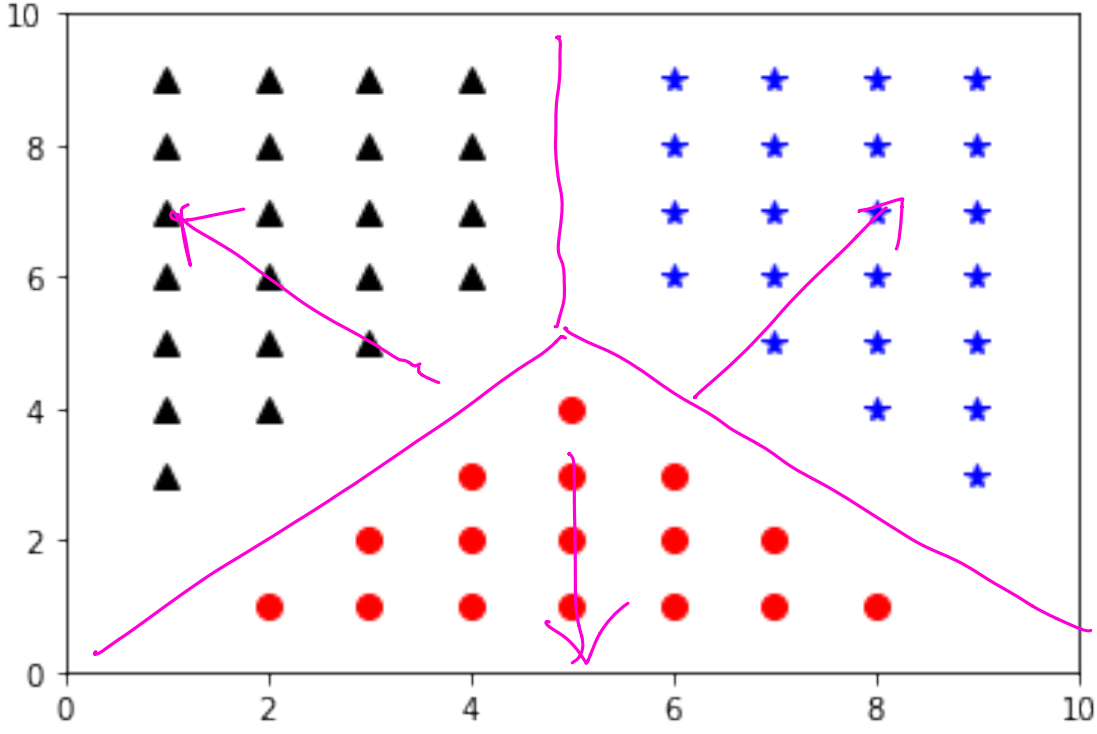
- Initialize $w_1 = \cdots = w_k = 0$ and $b_1 = \cdots = b_k = 0$
- Repeat while some training point (x, y) is misclassified:

for correct label y: $w_y = w_y + x$, $b_y = b_y + 1$

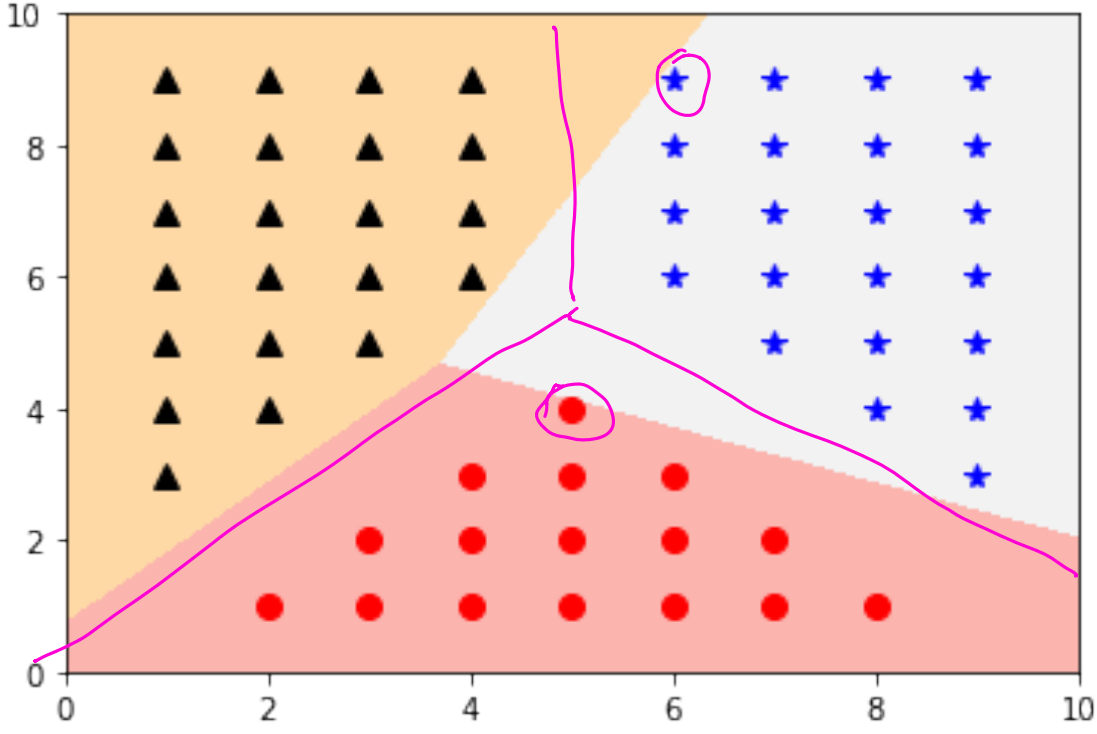
for predicted label $\widehat{y}$: $w_{\widehat{y}} = w_{\widehat{y}} - x$, $b_{\widehat{y}} = b_{\widehat{y}} - 1$



Multiclass Perceptron: example



Multiclass Perceptron: example



Multiclass SVM

Model: $w_1, \ldots, w_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$

Prediction: On instance x, predict label $\arg\max_j (w_j \cdot x + b_j)$

Learning. Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\min_{w_1,\ldots,w_k \in \mathbb{R}^d, b_1,\ldots,b_k \in \mathbb{R}, \xi \in \mathbb{R}^n} \sum_{j=1}^{k} \|w_j\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$w_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} - w_y \cdot x^{(i)} - b_y \geq 1 - \xi_i \quad \text{for all } i, \text{ all } y \neq y^{(i)}$$

$$\xi \geq 0$$

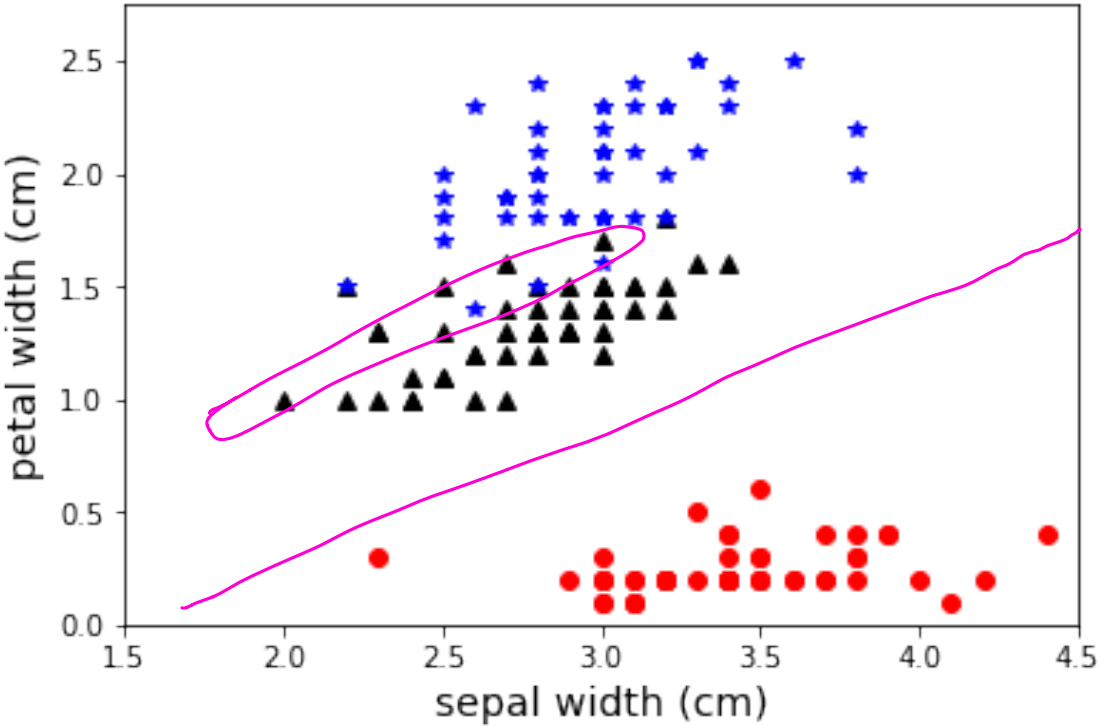
margin

slack variables

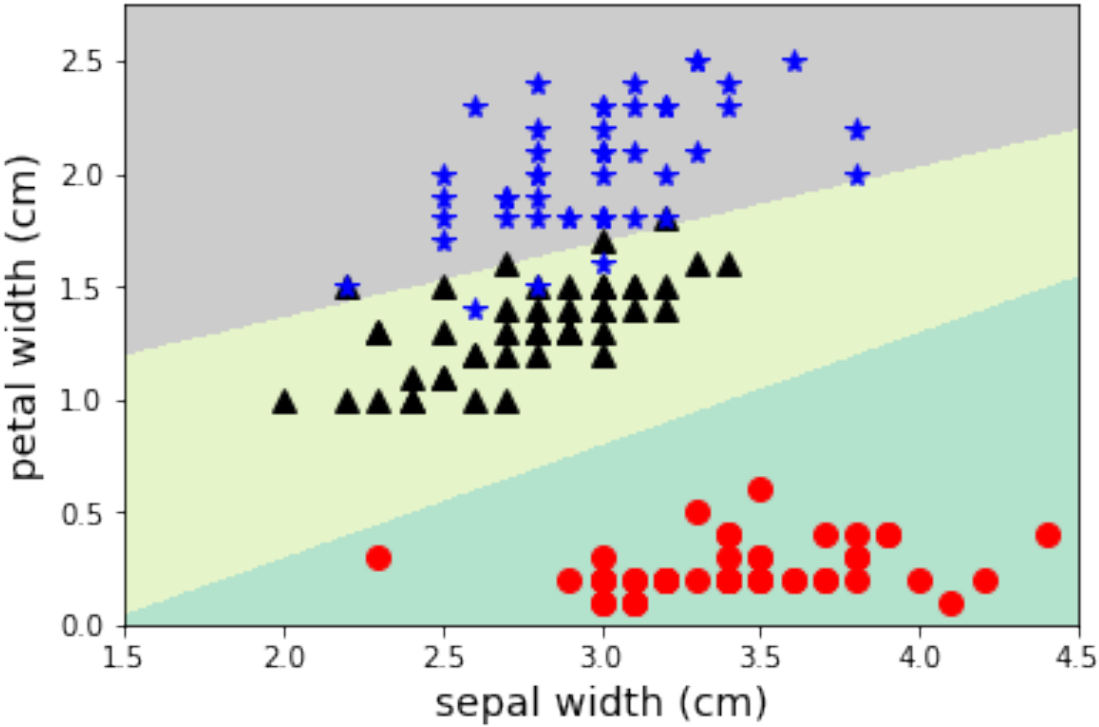
right label

wrong label

Multiclass SVM example: iris



Multiclass SVM example: iris



Multiclass SVM

Given training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\min_{w_1,\ldots,w_k \in \mathbb{R}^d, b_1,\ldots,b_k \in \mathbb{R}, \xi \in \mathbb{R}^n} \sum_{j=1}^{k} \|w_j\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$w_{y^{(i)}} \cdot x^{(i)} + b_{y^{(i)}} - w_y \cdot x^{(i)} - b_y \geq 1 - \xi_i \quad \text{for all } i, \text{ all } y \neq y^{(i)}.$$

$$\xi \geq 0 \qquad \rightarrow n$$

Once again, a convex optimization problem.

Question: how many variables and constraints do we have?

$k \times d + k + n = k(d+1) + n$ variables

$n + n(k-1)$ constraints