SVMs and Kernels

Two different approaches to regression/classification

Generative Approach

$x \sim P$ $\quad y = w^T x + \varepsilon,\ \varepsilon \sim N(0, I)$

$p(y|x) = \frac{1}{1+\exp(-w^T x)}$

- Assume something about P(x,y)
- Find f which maximizes likelihood of training data | assumption $\{(x_i, y_i)\}_{i=1}^n$ w maximize $\prod_{i=1}^n P(x_i, y_i | w)$
  - Often reformulated as minimizing loss

$\Leftrightarrow \min -\sum_{i=1}^n \log P(x_i, y_i | w)$

Versus

Discriminative Approach

l_2: regression

0/1: classification

linear function $\begin{cases} f(x) = w^T x \\ w^T x + b > 0 \Rightarrow 1 \\ w^T x + b < 0 \Rightarrow 0 \end{cases}$

- Pick a loss function
- Pick a set of hypotheses H
- Pick f from H which minimizes loss on training data

$$\min_{f \in H} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$$

☆ No assumption on $P(y|x)$

Our description of logistic regression was the former

- **Learn**: f:**X —>**Y
  - **X** – features
  - **Y – target classes**

$$Y \in \{-1, 1\}$$

- **Loss function:**

$$\ell(f(x), y) = \mathbb{1}\{f(x) \neq y\}$$

- **Expected loss of f:**

$$\mathbb{E}_{XY}[\mathbb{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbb{1}\{f(x) \neq Y\} \mid X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbb{1}\{f(x) \neq Y\} \mid X = x] = 1 - P(Y = f(x) \mid X = x)$$

- **Bayes optimal classifier:**

$$f(x) = \arg\max_y \underbrace{P(Y = y \mid X = x)}_{\text{need a model}}$$

- **Model of logistic regression:**

$$P(Y = y \mid x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

Our description of logistic regression was the former

- **Learn**: f:**X —>**Y
  - **X** – features
  - **Y – target classes**

$$Y \in \{-1, 1\}$$

- **Expected loss of f:**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = 1 - P(Y = f(x)|X = x)$$

- **Bayes optimal classifier:**

$$f(x) = \arg\max_y \mathbb{P}(Y = y|X = x)$$

- **Model of logistic regression:**

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

Generative

Discriminative

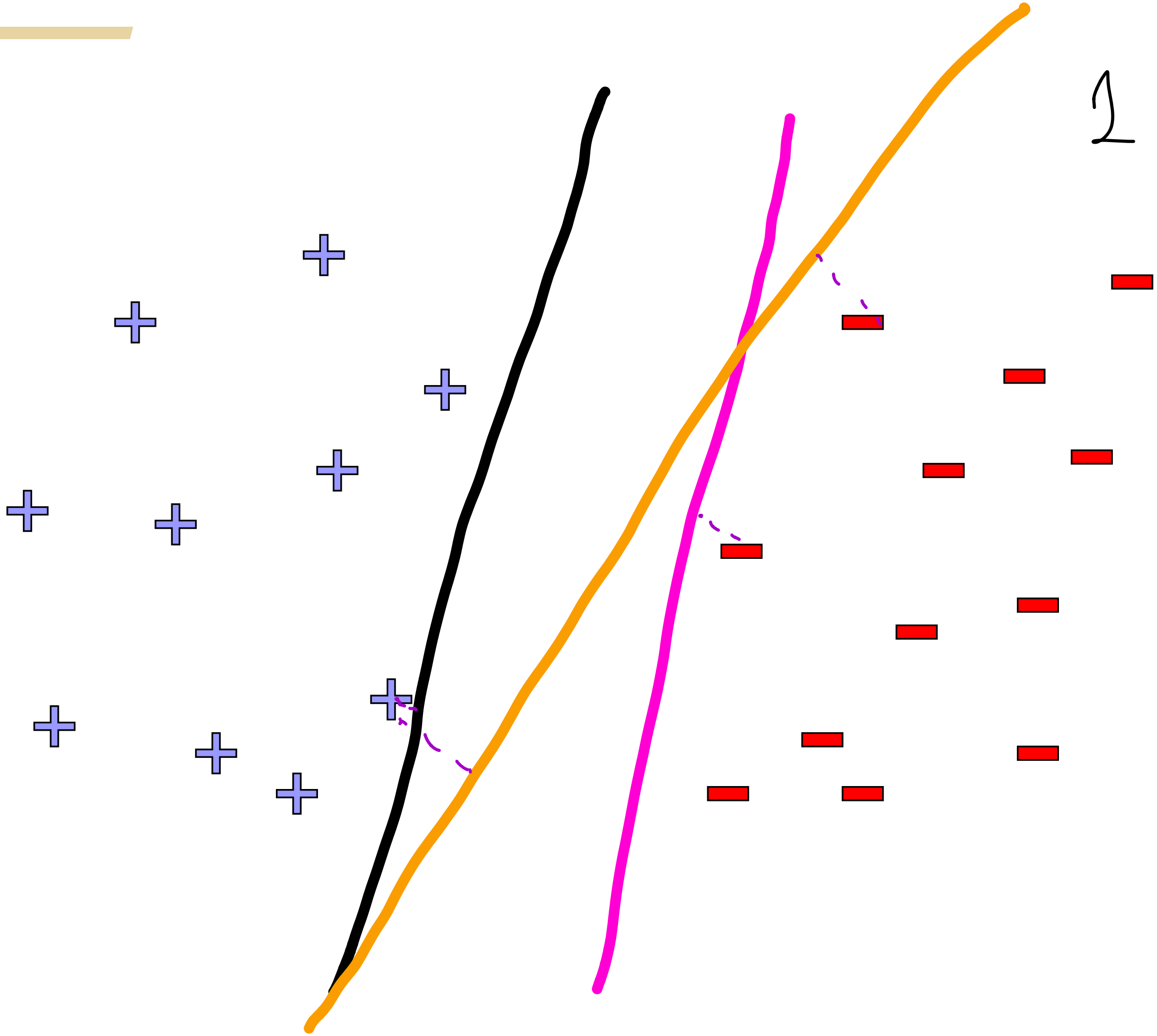
- **Loss function:**

$$\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$$

What if the model is wrong? What other ways can we pick linear decision rules?

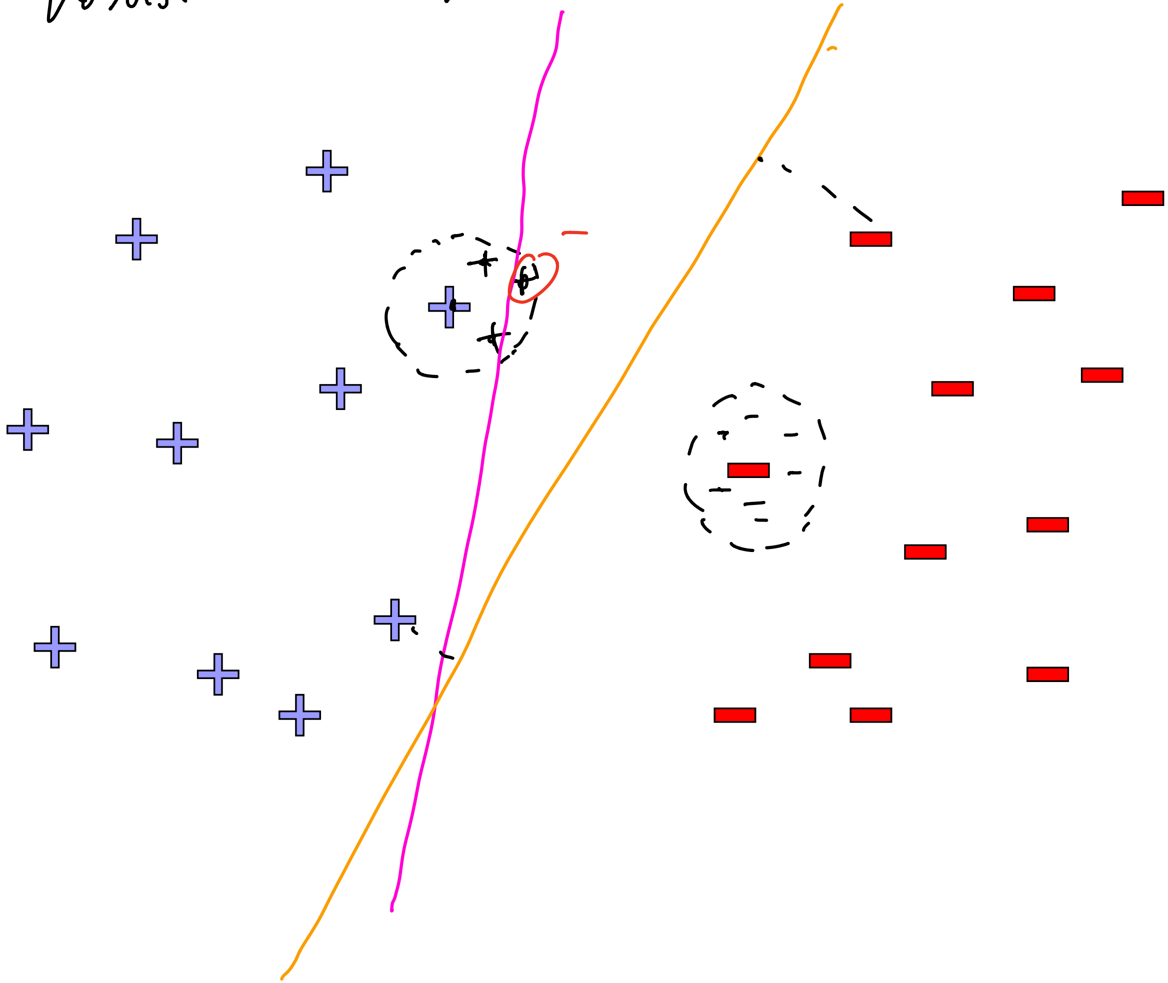
Linear classifiers – Which line is better?

$\mathbb{1}\{f(x) \neq y\}$



Linear classifiers – Which line is better?

robustness to perturbation



Linear classifiers – Which line is better?