Which classifier should you choose?

Depends on goal:

- Inference: logistic/probit regression, discriminant analysis
- prediction: discriminant analysis, KNN, SVM

SVMs with more than 2 classes

Difficult to extend SVMs to $K > 2$ classes.

Suppose $Y \in \{1, 2, \ldots, K\}$.

1 vs. 1 classification

Fit $\binom{K}{2}$ SVMs for comparing all pairs.

To predict at a new set of features, predict under all $\binom{K}{2}$ models, choose the class w/ highest count

1 vs all classification

Fit K SVMs for comparing ith class against everything else $\{1, 2, \dots, i-1, i+1, \dots, K\}$ → categorized as single class "not i", do for $i = 1, 2, \dots, K$.

For a new feature, look at $f_i(x^*)$, choose the class i for which this is biggest.

$*$ new feature

SVMs via penalization

Turns out that the SVM is the minimizer of

$$\sum_{i=1}^{n} \left(1 - y_i f(\underline{x}_i)\right)_+ + \lambda \|\beta\|_2^2$$

$$\sum_{j=1}^{p} \beta_j^2$$

where $f(\underline{x}) = \beta_0 + \sum_{j=1}^{n} \alpha_j y_j K(\underline{x}_j, \underline{x})$

and

$$(x)_+ = \begin{cases} x, & x > 0 \\ 0, & x \le 0 \end{cases}$$ hinge loss

Is in format $\sum_{i=1}^{n} L(y_i, \underline{x}_i, \beta_0, \underline{\beta}) + \lambda P(\underline{\beta})$

"Loss" "penalty"

where loss function L measures closeness of model to data + penalty controls the size of model.

"complexity"

Sanity check that $(x)_+$ is doing the right thing

$\hat{y} = \text{sign}(f)$

suppose $\underbrace{f \gg 0}$

really confident $\hat{y} = +1$

$$\begin{cases} \text{true } y = +1 & (1 - yf)_+ = 0 \quad \text{no loss} \\ \text{true } y = -1 & (1 - yf)_+ = 1 + f \gg 0 \quad \text{big loss} \end{cases}$$

Logistic regression $Y_i \in \{0,1\}$ $\underline{x}_1, \dots, \underline{x}_n$

model:
$$p(Y=1 \mid \underline{x}) = \frac{e^{f(\underline{x})}}{1+e^{f(\underline{x})}}$$

where $f(\underline{x}) = \beta_0 + \underline{\beta}^T \underline{x}$

the pmf for Y is $[f = f(\underline{x})]$

$$p(Y) = \left(\frac{e^f}{1+e^f}\right)^Y \left(1 - \frac{e^f}{1+e^f}\right)^{1-Y}, \quad Y \in \{0,1\}$$

$$= \left(\frac{e^f}{1+e^f}\right)^Y \left(\frac{1}{1+e^f}\right)^{1-Y}, \quad Y \in \{0,1\}$$

$$= \left(\frac{1}{1+e^{-f}}\right)^Y \left(\frac{1}{1+e^f}\right)^{1-Y}, \quad Y \in \{0,1\}$$

$= $ [recode for $y \in \{-1,+1\}$]

$$= \left(\frac{1}{1+e^{-f}}\right)^{\mathbb{1}[y=+1]} \left(\frac{1}{1+e^{f}}\right)^{\mathbb{1}[y=-1]}, \quad y \in \{-1,+1\}$$

$$-\log p(y) = \mathbb{1}_{[y=+1]} \log(1+e^{-f}) + \mathbb{1}_{[y=-1]} \log(1+e^{f})$$

$$= \text{claim} = \log(1+e^{-yf})$$

Check $y=+1 \Rightarrow \log(1+e^{-f})$ ✓

$y=-1 \Rightarrow \log(1+e^{f})$ ✓

In logistic regression, $\beta_0, \underline{\beta}$ are estimated by maximum likelihood, i.e. β_0 + $\underline{\beta}$ minimize

$$- \log p(y_1, \dots, y_n) \qquad \left[\text{Assume } y_i \in \{+1, -1\}\right]$$

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

$$= \sum_{i=1}^{n} -\log p(y_i)$$

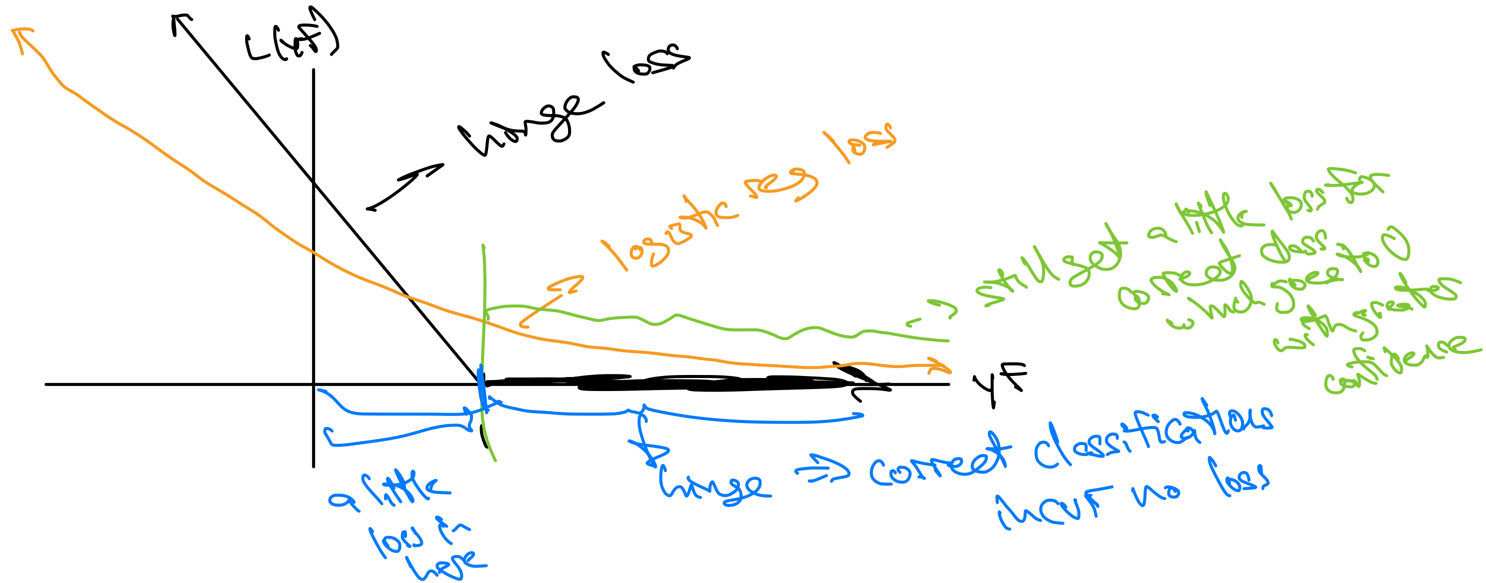
$$= \sum_{i=1}^{n} \log\left(1 + e^{-y_i f(\underline{x}_i)}\right) \quad \rightarrow \text{we minimize}$$

looks like $= \sum_{i=1}^{n} L(y_i, \underline{x}_i, \beta_0, \underline{\beta}) + \lambda P(\underline{\beta})$

in this case $P(\underline{\beta}) = 0$

* Summarize:

	SVMs	Logistic
Penalty	$\|\beta\|_2^2$	0
Loss fcn	$(1-yf)_+$	$\log(1+e^{-yf})$



Remark In logistic regression $\hat{\beta}_i$ never 0

In SVMs " are often 0

Is there a probalistic interpretation for SVMs?

$$p(Y=1 \mid x) \propto e^{-(1-yf)_+} \qquad Y \in \{-1,+1\}$$