CSE 575
Statistical Machine Learning

Lecture 8

YooJung Choi

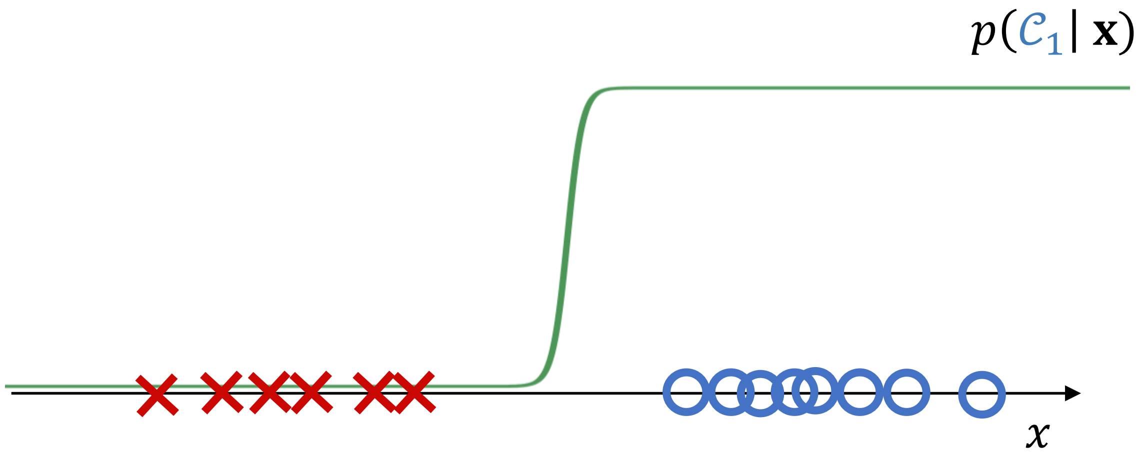
Fall 2022

Announcements

- Reminder: project proposal due this Friday, 9/23
- Homework will be posted tonight (9/19)
  - Due next Tuesday, 9/27
  - Solutions must be typed and submitted as PDF files
- Midterm 1: in-class, Wednesday 10/5

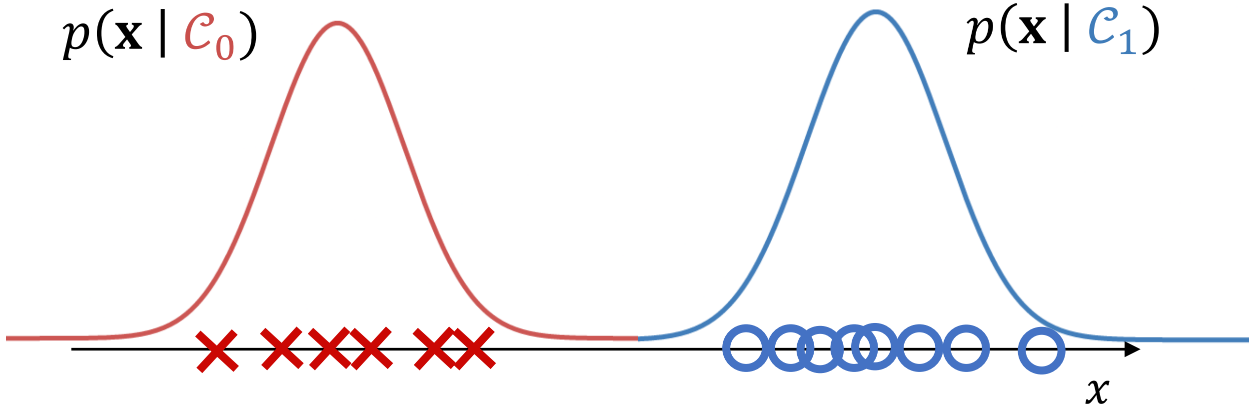
Probabilistic models for classification

- Discriminative models: $p(\mathcal{C}_k|\mathbf{x})$
- Generative models: $p(\mathbf{x}, \mathcal{C}_k)$
  - Often by modeling the class-conditional $p(\mathbf{x} \mid \mathcal{C}_k)$ and the class prior $p(\mathcal{C}_k)$
  - Use Bayes' theorem to compute $p(\mathcal{C}_k|\mathbf{x})$



Probabilistic models for classification

- Discriminative models: $p(\mathcal{C}_k|\mathbf{x})$
- Generative models: $p(\mathbf{x}, \mathcal{C}_k)$
  - Often by modeling the class-conditional $p(\mathbf{x} \mid \mathcal{C}_k)$ and the class prior $p(\mathcal{C}_k)$
  - Use Bayes' theorem to compute $p(\mathcal{C}_k|\mathbf{x})$



Bayes classifier

- How to make classifications given $p(\mathcal{C}_k|\mathbf{x})$?
- Misclassification probability (“*risk*”) of a classifier $y(\mathbf{x})$ on an example $\mathbf{x}$ associated with class $\mathcal{C}_k$: $P(y(\mathbf{x}) \neq \mathcal{C}_k)$ => want to minimize this risk
- Achieved by the *Bayes classifier*: $y(\mathbf{x}) = \mathrm{argmax}_k \, p(\mathcal{C}_k|\mathbf{x})$
- E.g. for binary class $t \in \{0,1\}$, the *decision function* of the Bayes classifier is:

$$y(\mathbf{x}) = \begin{cases} 1 & \text{if } p(t = 1|\mathbf{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

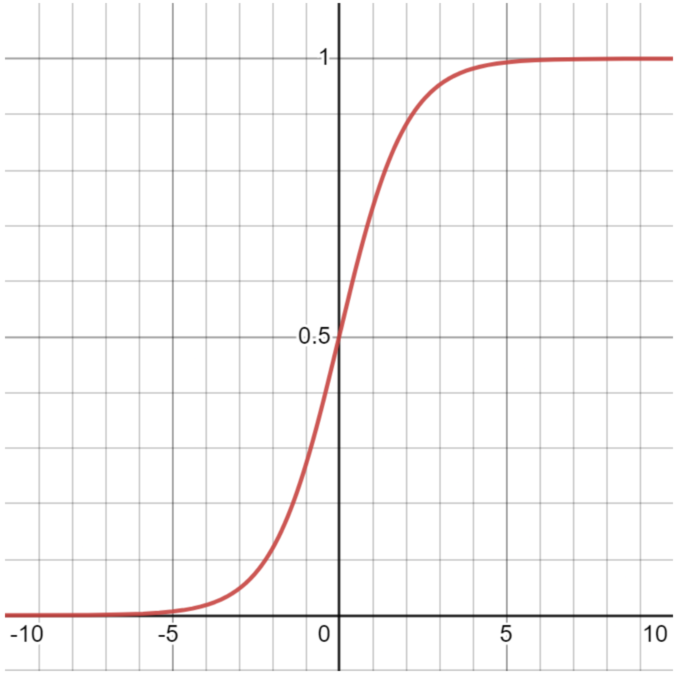
Posterior probability of class

- Using Bayes' theorem:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

Logistic sigmoid function



where $a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$

- Note: $\sigma(a) \geq 0.5$ iff $a \geq 0$
- i.e., decision boundary $\{\mathbf{x}: \sigma(a) = 0.5\} = \{\mathbf{x}: a = 0\}$

Posterior probability of class

- (*Multi-class case*) Using Bayes' theorem:

softmax function

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)} = s(a_k)$$

where $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$

- Intuitively, a smooth version of max:

If $a_k \gg a_j$ for all $j \neq k$, $p(\mathcal{C}_k|\mathbf{x}) \approx 1$ and $p(\mathcal{C}_j|\mathbf{x}) \approx 0$

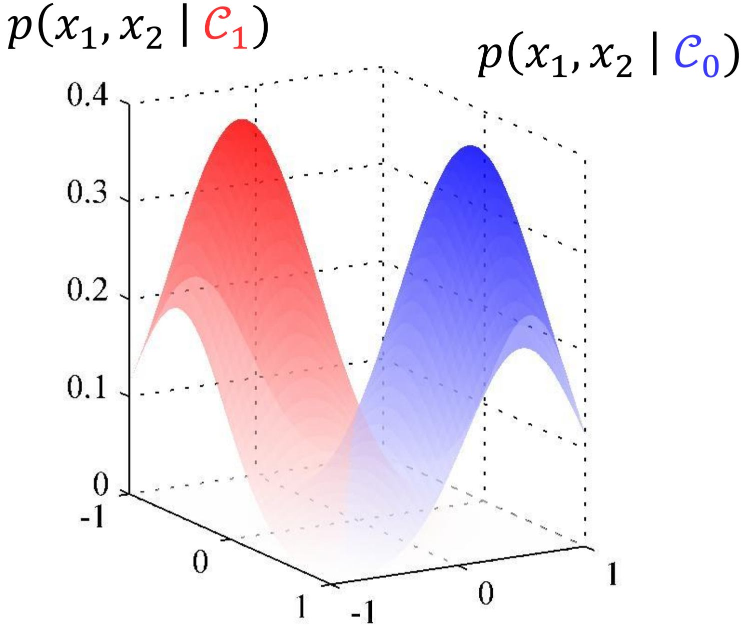
- The decision rule $\operatorname{argmax}_k p(\mathcal{C}_k|\mathbf{x})$ is equivalent to $\operatorname{argmax}_k a_k$

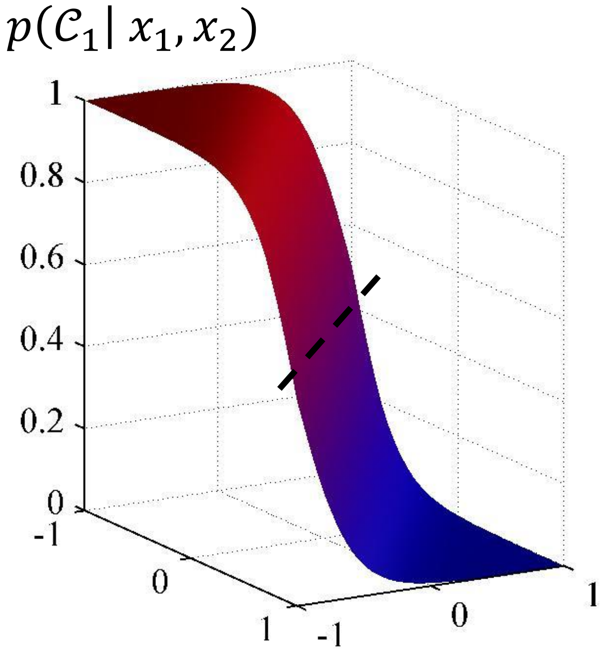
Gaussian discriminant analysis

- Consider D continuous features $\mathbf{x}$ and binary class $t \in \{0,1\}$
- Bernoulli class prior: $p(t) = \phi^t(1-\phi)^{1-t}$
- Let's assume the class conditional $p(\mathbf{x}|t)$ are multivariate Gaussians
- For now, also assume that the covariance matrix is the same between classes

$$p(\mathbf{x}|t) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_t)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_t)\right\}$$

Gaussian discriminant analysis





Gaussian discriminant analysis

- Class prior: $p(t) = \phi^t (1-\phi)^{1-t}$
- Class conditional: $p(\mathbf{x}|t) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_t)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_t)\right\}$
- Posterior probability $p(t=1|\mathbf{x}) = \sigma(a)$ where:

$$a = \ln \frac{p(\mathbf{x}|t=1)p(t=1)}{p(\mathbf{x}|t=0)p(t=0)} = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + \ln\frac{\phi}{1-\phi}$$

$$= \mathbf{w}^T\mathbf{x} + w_0$$

linear decision boundary!

where $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, $w_0 = -\frac{1}{2}\boldsymbol{\mu}_1{}^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0{}^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_0 + \ln\frac{\phi}{1-\phi}$

- => Linear discriminant analysis (LDA)
- If classes do not share the covariance matrix => quadratic discriminant analysis (QDA)

Gaussian discriminant analysis

- Parameters: $\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}$
- Log-likelihood given N examples:

$$ll(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n | t_n) \cdot p(t_n)$$

$$= \ln \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{t_n}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{t_n}\right)\right\} \cdot \phi^{t_n}(1-\phi)^{1-t_n}$$

$$= \sum_{n=1}^{N} \left(-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{t_n}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{t_n}\right) + t_n \ln\phi + (1-t_n)\ln(1-\phi)\right)$$

Gaussian discriminant analysis

Maximum-likelihood estimates:

$$\phi = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} t_n = \frac{N_1}{N_1 + N_0} \text{ where } N_1 \text{ is the number of examples s.t. } t_n = 1$$

$$\boldsymbol{\mu}_1 = \frac{1}{\mathrm{N}_1} \sum_{n=1}^{N} t_n \, \mathbf{x}_n, \qquad \boldsymbol{\mu}_0 = \frac{1}{\mathrm{N}_0} \sum_{n=1}^{N} (1 - t_n) \, \mathbf{x}_n$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}_{t_n}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_{t_n}\right)^T = \frac{N_1}{N} \boldsymbol{\Sigma}_1 + \frac{N_0}{N} \boldsymbol{\Sigma}_0$$

$$\text{where } \boldsymbol{\Sigma}_1 = \frac{1}{\mathrm{N}_1} \sum_{n:t_n=1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T$$

Gaussian discriminant analysis

- Can be easily extended to *K > 2* case
- E.g. posterior probability $p(\mathcal{C}_k|\mathbf{x}) = s(a_k)$ where:

$$a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k) = \mathbf{w}_k^T\mathbf{x} + w_{k0}$$

where $\mathbf{w} = \mathbf{\Sigma}^{-1}\,\boldsymbol{\mu}_k,\; w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k{}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$

- Again, a linear discriminant

Generative learning: discrete data

- Consider D *discrete* features $\mathbf{x}$ and binary class $t \in \{0,1\}$
- If the features are binary, there are 2^D possible instantiations of the features
- $2^D - 1$ independent parameters for each class conditional $p(\mathbf{x}|t)$
- Too expensive!

Naïve Bayes

- Naïve Bayes assumption: *features are independent given class*

$$P(X_1, X_2|C) = P(X_1|X_2, C) \cdot P(X_2, C) \qquad \text{Product rule}$$
$$= P(X_1|C) \cdot P(X_2|C)$$

- For D features, $P(X_1, \dots, X_D\,|C) = \prod_{i=1}^{D} P(X_i|C)$
- E.g. $P(\text{BloodTest}, \text{UrineTest} \mid \text{Pregnant}t) = P(\text{BloodTest} \mid \text{Pregnant}t) \times P(\text{UrineTest} \mid \text{Pregnant}t)$
- D independent parameters to represent each class conditional $P(X_1, \dots, X_D\,|C)$

Naïve Bayes

- Class prior: $p(t) = \phi^t(1-\phi)^{1-t}$
- Class conditional: $p(\mathbf{x}|t) = \prod_{i=1}^{D} p(x_i|t) = \prod_{i=1}^{D} \mu_{ti}^{x_i}(1-\mu_{ti})^{1-x_i}$
- Parameters: $\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1$
- Log-likelihood given *N* examples:

$$ll(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n|t_n) \cdot p(t_n)$$

$$= \sum_{n=1}^{N} \left(\sum_{i=1}^{D} \left(x_{ni} \ln \mu_{t_n i} + (1 - x_{ni}) \ln\left(1 - \mu_{t_n i}\right)\right) + t_n \ln \phi + (1 - t_n) \ln(1-\phi) \right)$$

Naïve Bayes

- Maximum-likelihood estimates:

$$\phi = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} t_n = \frac{N_1}{N_1 + N_0} \text{ where } N_1 \text{ is the number of examples s.t. } t_n = 1$$

$$\boldsymbol{\mu}_1 = \frac{1}{\mathrm{N}_1} \sum_{n=1}^{N} t_n \mathbf{x}_n, \qquad \boldsymbol{\mu}_0 = \frac{1}{\mathrm{N}_0} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n$$

	x_1 UrineTest?	x_2 BloodTest?	t Pregnant?
	1	0	0
	0	1	1
→	1	1	1
→	1	0	1
	⋮	⋮	⋮
→	0	1	0

$$\phi = p(t = 1) = \frac{N_1}{N}$$

$$\mu_{11} = p(x_1 = 1 \mid t = 1) = \frac{\#\{x_1 = 1, t = 1\}}{N_1}$$

$$\mu_{02} = p(x_2 = 1 \mid t = 0) = \frac{\#\{x_2 = 1, t = 0\}}{N_0}$$

$$\vdots$$

Naïve Bayes

- Class prior: $p(t) = \phi^t(1-\phi)^{1-t}$
- Class conditional: $p(\mathbf{x}|t) = \prod_{i=1}^{D} p(x_i|t) = \prod_{i=1}^{D} \mu_{ti}^{x_i}(1-\mu_{ti})^{1-x_i}$
- Posterior probability

$$p(t=1|\mathbf{x}) = \frac{\prod_{i=1}^{D} p(x_i|t=1)\, p(t=1)}{\prod_{i=1}^{D} p(x_i|t=1)\, p(t=1) + \prod_{i=1}^{D} p(x_i|t=0)\, p(t=0)}$$

- Alternatively, $p(t=1|\mathbf{x}) = \sigma(a)$ where:

$$a = \ln\frac{p(\mathbf{x}|t=1)p(t=1)}{p(\mathbf{x}|t=0)p(t=0)} = \ln\prod_{i=1}^{D}\frac{\mu_{1i}^{x_i}(1-\mu_{1i})^{1-x_i}}{\mu_{0i}^{x_i}(1-\mu_{0i})^{1-x_i}} + \ln\frac{\phi}{1-\phi}$$

$$= \sum_{i=1}^{D} x_{\mathrm{i}}\ln\frac{\mu_{1i}(1-\mu_{0i})}{\mu_{0i}(1-\mu_{1i})} + \sum_{i=1}^{D}\ln\frac{(1-\mu_{1i})}{(1-\mu_{0i})} + \ln\frac{\phi}{1-\phi}$$

linear decision boundary!

Observation

- Bernoulli class prior + Gaussian class-conditional => class posterior looks like $\sigma(\mathbf{w}^T\mathbf{x} + w_0)$
- Bernoulli class prior + (categorical) naïve Bayes class-conditional => $\sigma(\mathbf{w}^T\mathbf{x} + w_0)$
- Exponential family as class-conditional => generalized linear model $\sigma(\mathbf{w}^T\mathbf{x} + w_0)$ or $s(\mathbf{w}^T\mathbf{x} + w_0)$ as the class posterior

  e.g. Gaussian, Bernoulli, categorical, Poisson, Beta, Dirichlet, …

- What if we learn the class posterior probability $p(\mathcal{C}_k|\mathbf{x})$ as $\sigma(\mathbf{w}^T\mathbf{x} + w_0)$ or $s(\mathbf{w}^T\mathbf{x} + w_0)$ directly?

Generative vs Discriminative

Generative models:

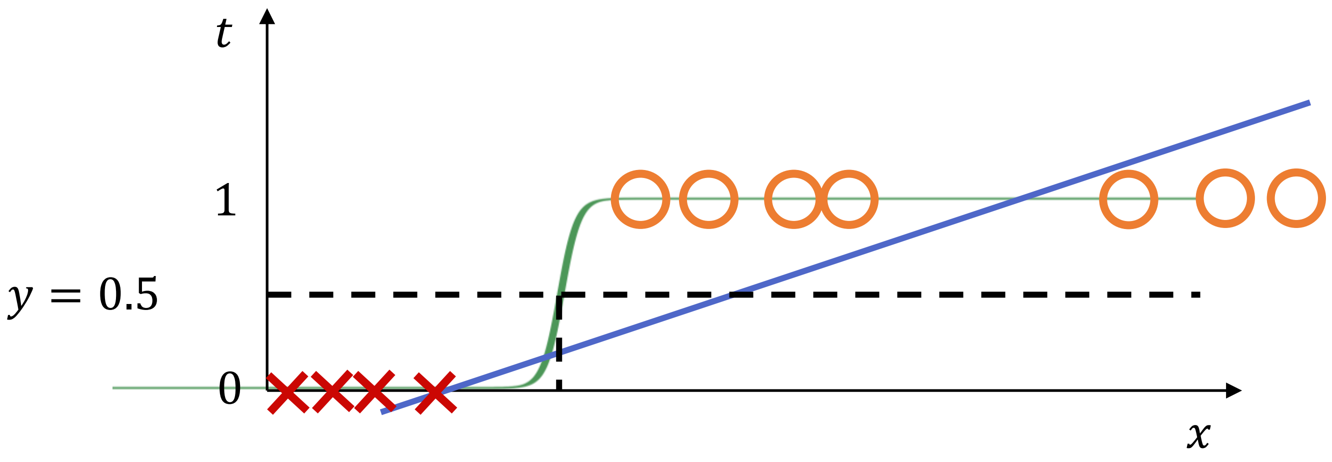
- Can be used for various tasks:
  - Sampling and generating synthetic data points
  - Outlier detection
  - Prediction with missing values
  - Many more probabilistic queries…
- Performs very well if the modeling assumptions hold
- Tend to have more parameters

Discriminative models:

- Only useful for classification
- “don’t solve a harder problem as an intermediate step”
- Tend to have fewer parameters

Logistic regression

- Model $p(t=1|\mathbf{x})$ via $\mathrm{y}(\mathbf{x}) = \frac{1}{1+\exp\{-\mathbf{w}^T\mathbf{x}\}} = \sigma(\mathbf{w}^T\mathbf{x})$
- Again, assume $x_0 = 1$
- Recall: linear regression failed on this example



Logistic regression

- Given *N* data points $\{(\mathbf{x}_n, t_n)\}$, the likelihood function is:

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n} \qquad \text{where } y_n = y(\mathbf{x}_n) = \sigma(\mathbf{w}^T\mathbf{x}_n)$$

- Maximize log-likelihood, or equivalently, minimize the negative log-likelihood as the error function:

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\}$$