Announcements

Assignments

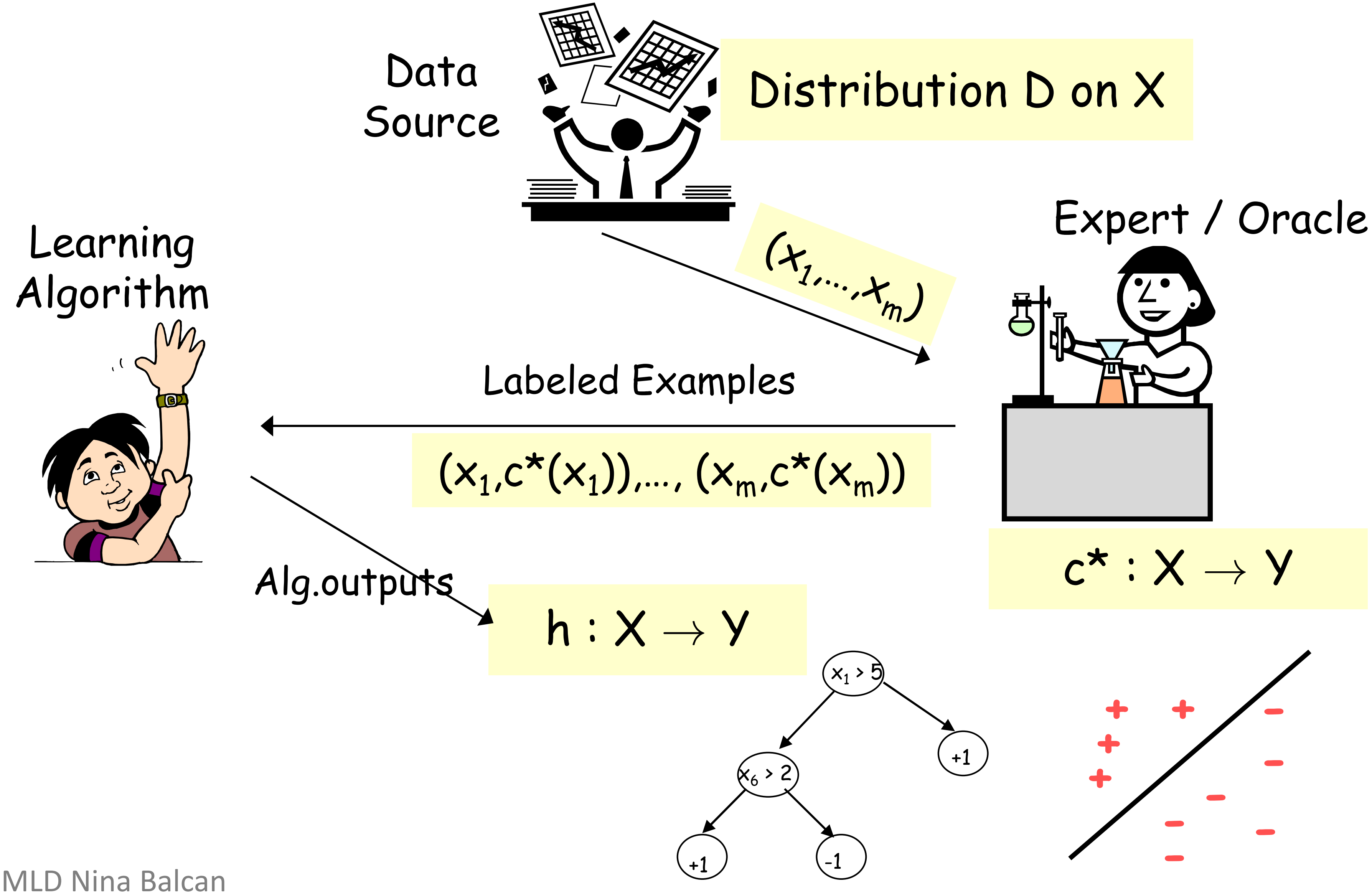
- HW10 (programming + “written”)
  - Due Thu 4/30, 11:59 pm

Introduction to Machine Learning

Learning Theory

Instructor: Pat Virtue

Model for Supervised Learning



Slide credit: CMU MLD Nina Balcan

Optimal Classification Function

Find the best $h(x) \to \hat{y}$ by searching in the space of hypothesis functions $h \in \mathcal{H}$.

Optimal classifier:

$$h^*(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y \mid X = x)$$

But why?

Optimal Decision Boundaries

Decision boundary

- The set of points in the domain of the input (x) where the predicted classification changes

Two class decision boundary

- So far, we have decided to let the decision boundary be all x such that:

$$p(Y = 0 \mid X = x) = p(Y = 1 \mid X = x)$$

- What assumptions are we making here?
    - This assumes that the cost of predicting it wrong is the same for both classes

Optimal Classification Function

Find the best $h(x) \rightarrow \hat{y}$ by searching in the space of hypothesis functions $h \in \mathcal{H}$.

Optimal classifier:

$$h^*(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y \mid X = x)$$

But why?

Goal: find a prediction function $h^*: \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes the expected loss for randomly drawn test data (X, Y)

$$h^* = \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

$L(y, \hat{y})$ is the loss or cost of predicting $\hat{y}$ when the true value is y.

Loss Functions

$$h^* = \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\left[L\left(Y, h(X)\right)\right]$$

Loss function:

$L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Classification:

- Two-class, 0,1 loss
- Two-class, arbitrary loss

False positives and false negatives:

Loss Functions

$$h^* = \operatorname*{argmin}_h \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

Loss function:

$L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Classification:

- Two-class, 0,1 loss

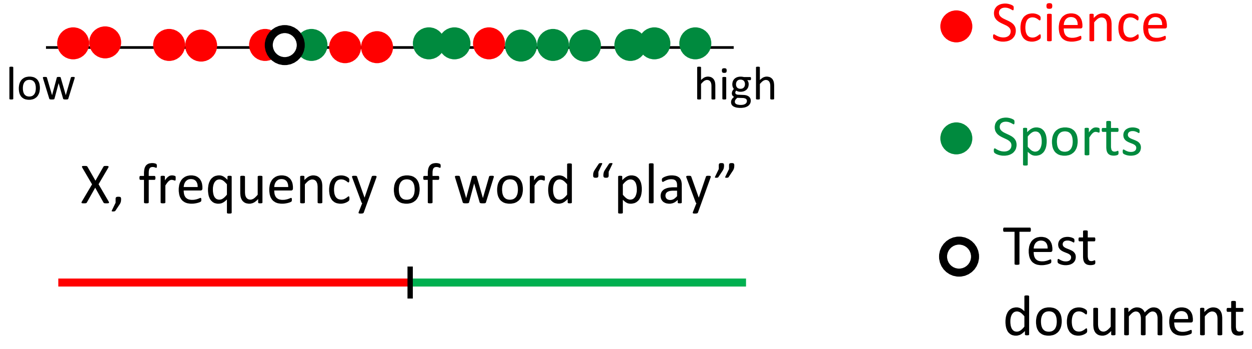
- Two-class, arbitrary loss

Regression:

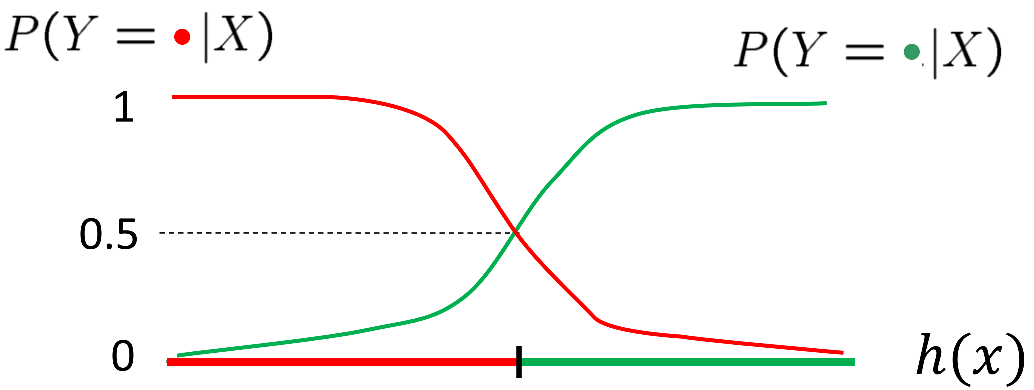
Expected Value

Quick review

Binary Classification



Model X and Y as random variables



For a given x, $h(x)$ = label Y which is more likely

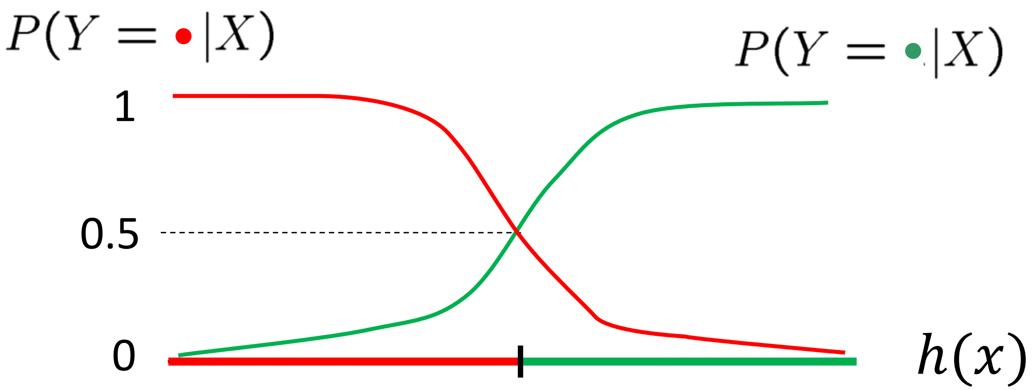
$$h(x) = \arg\max_{Y=y} P(Y = y | X = x)$$

Optimal Classification Function

$$h^*(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} P(Y = y \mid X = x)$$

$$h^* = \underset{h}{\operatorname{argmin}} \mathbb{E}_{XY}\left[L\left(Y, h(X)\right)\right]$$

Start with arbitrary two-class loss $L(y, \hat{y})$



Optimal Classification Function

Expected loss is also called risk:

$$R(h) = \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

$$h^* = \underset{h}{\operatorname{argmin}}\, R(h)$$

$$= \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

{Whiteboard derivation}

Optimal Classification Function

$$h^*(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} P(Y = y \mid X = x)$$

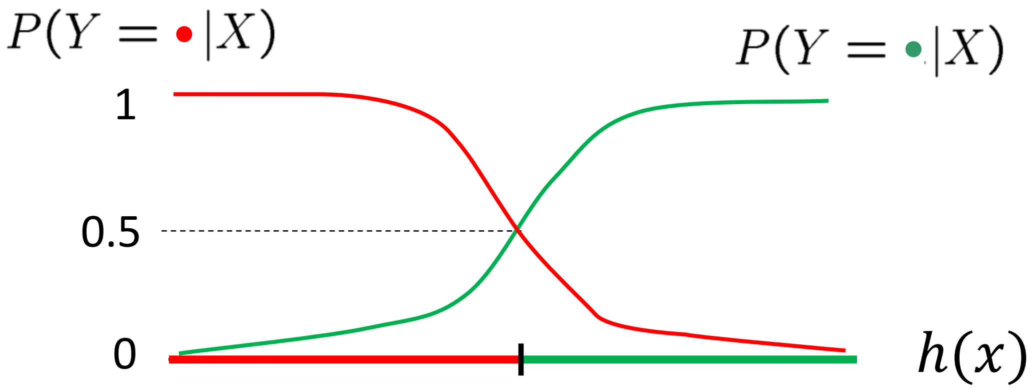
$$h^* = \underset{h}{\operatorname{argmin}} \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

Start with arbitrary two-class loss $L(y, \hat{y})$

$$h^*(x) = \begin{cases} 1 & \text{if} \quad P(Y = 0 \mid x)L(0,1) + P(Y = 1 \mid x)L(1,1) \\ & \qquad \le P(Y = 0 \mid x)L(0,0) + P(Y = 1 \mid x)L(1,0) \\ 0 & \text{otherwise} \end{cases}$$

Two-class, 0, 1 loss

$$h^*(x) = \begin{cases} 1 & \text{if} \quad P(Y = 0 \mid x) \\ & \qquad \le P(Y = 1 \mid x) \\ 0 & \text{otherwise} \end{cases}$$



Optimal Classification Function

$$h^*(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, P(Y = y \mid X = x)$$

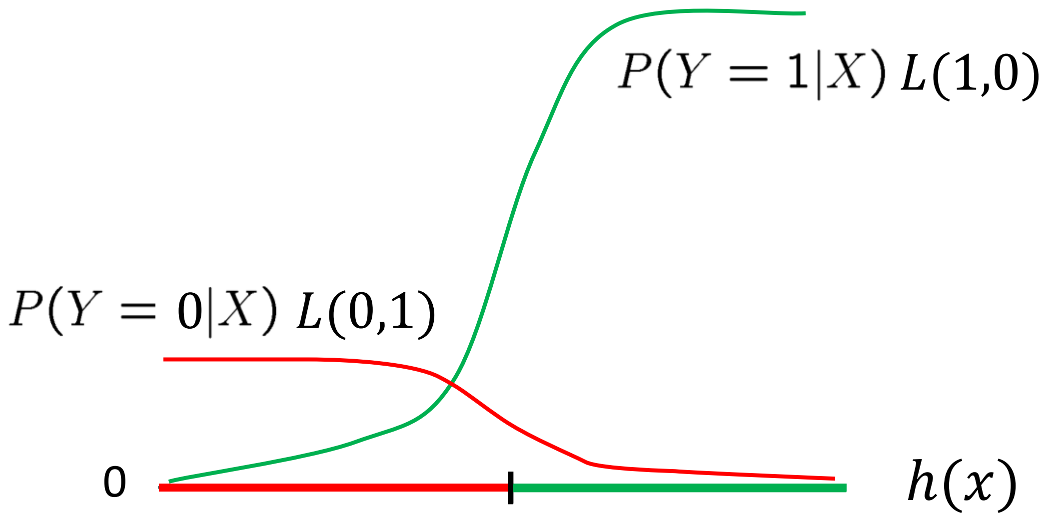
$$h^* = \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\big[L\big(Y, h(X)\big)\big]$$

Start with arbitrary two-class loss $L(y, \hat{y})$

$$h^*(x) = \begin{cases} 1 \text{ if} & P(Y=0 \mid x)L(0,1) + P(Y=1 \mid x)L(1,1) \\ & \leq P(Y=0 \mid x)L(0,0) + P(Y=1 \mid x)L(1,0) \\ 0 & \text{otherwise} \end{cases}$$

Two-class, weighted loss

$$h^*(x) = \begin{cases} 1 \text{ if} & P(Y=0 \mid x)L(0,1) \\ & \leq P(Y=1 \mid x)L(1,0) \\ 0 & \text{otherwise} \end{cases}$$

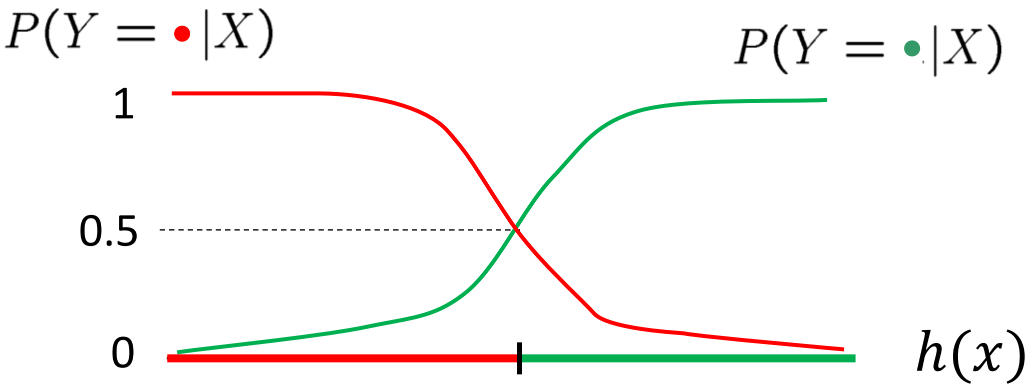


Optimal Classification Function

$$h^*(x) = \begin{cases} 1 & \text{if} \quad P(Y=0 \mid x) \leq P(Y=1 \mid x) \\ 0 & \text{otherwise} \end{cases}$$

What is the risk of the optimal classifer?

$$R(h^*) = \mathbb{E}_{XY}\left[L\big(Y, h^*(X)\big)\right]$$



Risk in Regression

Squared error loss $L(y, h(x)) = (h(x) - y)^2$

$$h^* = \underset{h}{\operatorname{argmin}} \, \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

Optimal Hypothesis Function

Goal: find a prediction function $h^*: \mathcal{X} \to \mathcal{Y}$ that minimizes the risk, the expected loss for randomly drawn test data (X, Y)

$$h^* = \underset{h}{\operatorname{argmin}}\, R(h) = \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\left[L\big(Y, h(X)\big)\right]$$

Learning from Training Data

But we want our hypothesis function to generalize well?

- How do we characterize and quantify this trade-off?
- {Back to the whiteboard}