CS189/CS289A
Introduction to Machine Learning
Lecture 4: Decision Theory

Peter Bartlett

January 29, 2015

Outline

Outline

- Decision theory

Outline

- Decision theory
    - Loss functions

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk.

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk.
  - Bayes decision rule.

Outline

- Decision theory
    - Loss functions
    - Probabilistic assumptions
    - Risk.
    - Bayes decision rule.
    - Excess risk.

Outline

- Decision theory
    - Loss functions
    - Probabilistic assumptions
    - Risk.
    - Bayes decision rule.
    - Excess risk.
    - Risk, Bayes decision rule, excess risk in regression.

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk.
  - Bayes decision rule.
  - Excess risk.
  - Risk, Bayes decision rule, excess risk in regression.
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules.

Outline

- Decision theory
  - **Loss functions**
  - Probabilistic assumptions
  - Risk.
  - Bayes decision rule.
  - Excess risk.
  - Risk, Bayes decision rule, excess risk in regression.
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules.

Loss Functions

The Prediction Problem

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

For example, the patterns $x \in \mathcal{X}$ might be vectors in $\mathbb{R}^{400}$.

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

For example, the patterns $x \in \mathcal{X}$ might be vectors in $\mathbb{R}^{400}$.
The labels $y \in \mathcal{Y}$ might be class labels in $\{0, 1, \ldots, 9\}$.

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

For example, the patterns $x \in \mathcal{X}$ might be vectors in $\mathbb{R}^{400}$.
The labels $y \in \mathcal{Y}$ might be class labels in $\{0, 1, \ldots, 9\}$.

To define the notion of a 'good prediction,' we can define a **loss function**

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

For example, the patterns $x \in \mathcal{X}$ might be vectors in $\mathbb{R}^{400}$.
The labels $y \in \mathcal{Y}$ might be class labels in $\{0, 1, \ldots, 9\}$.

To define the notion of a 'good prediction,' we can define a **loss function**

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

$\ell(\hat{y}, y)$ is the cost of predicting $\hat{y}$ when the outcome is y.

Loss Functions

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

For example, the patterns $x \in \mathcal{X}$ might be vectors in $\mathbb{R}^{400}$.
The labels $y \in \mathcal{Y}$ might be class labels in $\{0, 1, \ldots, 9\}$.

To define the notion of a 'good prediction,' we can define a **loss function**

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

$\ell(\hat{y}, y)$ is the cost of predicting $\hat{y}$ when the outcome is y.

Aim: $\ell(f(x), y)$ small.

Example: Classification

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$					
$y = 1$					
$y = 2$					
$\vdots$					
$y = 9$					

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0				
$y = 1$					
$y = 2$					
$\vdots$					
$y = 9$					

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0	1	1	$\cdots$	1
$y = 1$					
$y = 2$					
$\vdots$					
$y = 9$					

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0	1	1	$\cdots$	1
$y = 1$		0			
$y = 2$					
$\vdots$					
$y = 9$					

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0	1	1	$\cdots$	1
$y = 1$	1	0			
$y = 2$					
$\vdots$					
$y = 9$					

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0	1	1	$\cdots$	1
$y = 1$	1	0	1	$\cdots$	1
$y = 2$					
$\vdots$					
$y = 9$					

Loss Functions

Example: Classification

If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

$\ell(\hat{y}, y)$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 2$	$\cdots$	$\hat{y} = 9$
$y = 0$	0	1	1	$\cdots$	1
$y = 1$	1	0	1	$\cdots$	1
$y = 2$	1	1	0		1
$\vdots$	$\vdots$	$\vdots$		$\ddots$	$\vdots$
$y = 9$	1	1	1	$\cdots$	0

Example: Spam Classification

Example: Spam Classification

We might be more concerned about some errors than others:

Loss Functions

Example: Spam Classification

We might be more concerned about some errors than others:

$\ell(\hat{y}, y)$	$\hat{y} =$ Spam	$\hat{y} =$ Ham
$y =$ Spam		
$y =$ Ham		

Loss Functions

Example: Spam Classification

We might be more concerned about some errors than others:

$\ell(\hat{y}, y)$	$\hat{y}$ = Spam	$\hat{y}$ = Ham
y = Spam	0	
y = Ham		0

Loss Functions

Example: Spam Classification

We might be more concerned about some errors than others:

$\ell(\hat{y}, y)$	$\hat{y} =$ Spam	$\hat{y} =$ Ham
$y =$ Spam	0	1
$y =$ Ham		0

Loss Functions

Example: Spam Classification

We might be more concerned about some errors than others:

$\ell(\hat{y}, y)$	$\hat{y} = \text{Spam}$	$\hat{y} = \text{Ham}$
$y = \text{Spam}$	0	1
$y = \text{Ham}$	100	0

Loss Functions

Example: Regression

Loss Functions

Example: Regression

Outcomes are in $\mathcal{Y} = \mathbb{R}$

Loss Functions

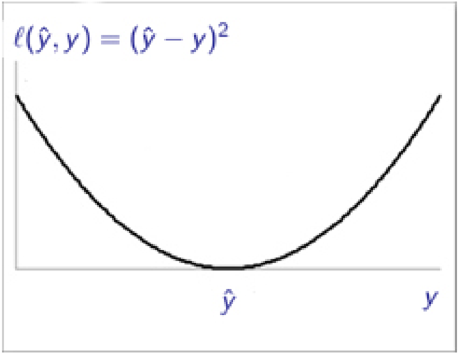
Example: Regression

Outcomes are in $\mathcal{Y} = \mathbb{R}$, we might choose the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

Loss Functions

Example: Regression

Outcomes are in $\mathcal{Y} = \mathbb{R}$, we might choose the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.



Outline

- Decision theory
  - Loss functions
  - **Probabilistic assumptions**
  - Risk.
  - Bayes decision rule.
  - Excess risk.
  - Risk, Bayes decision rule, excess risk in regression.
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules.

Probabilistic assumptions for prediction problems

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

- We need to assume something about the relationship between the data $(x_1, y_1), \ldots, (x_n, y_n)$ and the subsequent (x, y) pairs.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \dots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

- We need to assume something about the relationship between the data $(x_1, y_1), \dots, (x_n, y_n)$ and the subsequent (x, y) pairs.
- A common formulation: Assume that they are randomly chosen, and have the same probability distribution.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

- We need to assume something about the relationship between the data $(x_1, y_1), \ldots, (x_n, y_n)$ and the subsequent (x, y) pairs.
- A common formulation: Assume that they are randomly chosen, and have the same probability distribution.
- This is not an unreasonable assumption.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

- We need to assume something about the relationship between the data $(x_1, y_1), \ldots, (x_n, y_n)$ and the subsequent (x, y) pairs.
- A common formulation: Assume that they are randomly chosen, and have the same probability distribution.
- This is not an unreasonable assumption. But keep in mind that it is a model that is typically wrong at some level of detail.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (x, y) pairs, $f(x)$ is a good prediction of y.

- We need to assume something about the relationship between the data $(x_1, y_1), \ldots, (x_n, y_n)$ and the subsequent (x, y) pairs.
- A common formulation: Assume that they are randomly chosen, and have the same probability distribution.
- This is not an unreasonable assumption. But keep in mind that it is a model that is typically wrong at some level of detail.

  (Think about the MNIST digits data versus your handwriting.)

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (X, Y) pairs, $f(X)$ is a good prediction of Y.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (X, Y) pairs, $f(X)$ is a good prediction of Y.

1. Assume that (X_i, Y_i) and (X, Y) are chosen i.i.d. (independently and identically distributed), according to some probability distribution on $\mathcal{X} \times \mathcal{Y}$.

Detour: Joint and conditional distributions

Detour: Joint and conditional distributions

$P(X, Y)$

Detour: Joint and conditional distributions

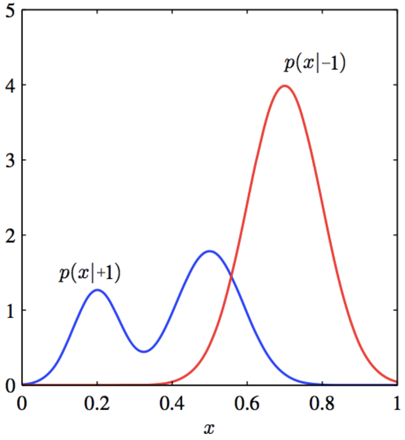
$$P(X, Y) = P(Y)P(X|Y)$$

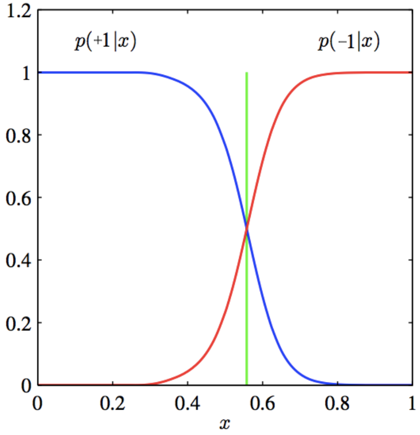
Detour: Joint and conditional distributions

$$P(X, Y) = P(Y)P(X|Y) = P(X)P(Y|X).$$

Detour: Joint and conditional distributions

$$P(X, Y) = P(Y)P(X|Y) = P(X)P(Y|X).$$

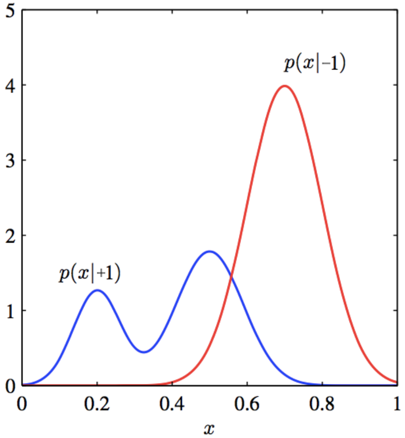


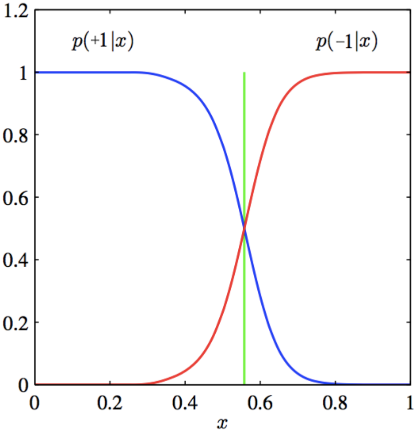


Detour: Joint and conditional distributions

$$P(X, Y) = P(Y)P(X|Y) = P(X)P(Y|X).$$

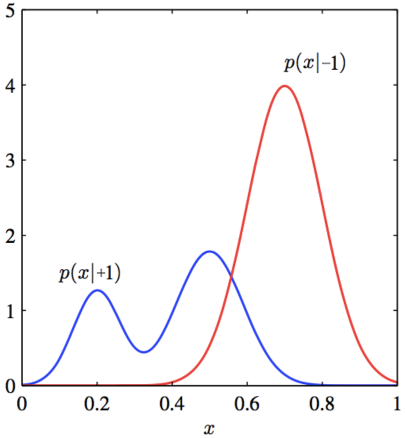
$P(X \in S \text{ and } Y = 1)$

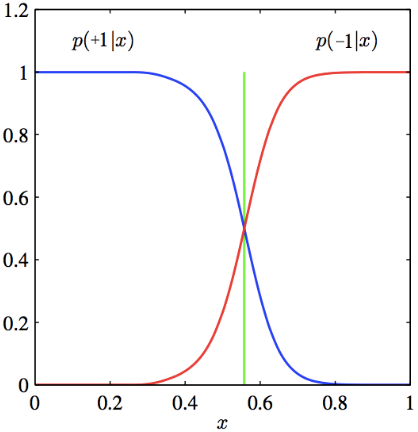




Detour: Joint and conditional distributions

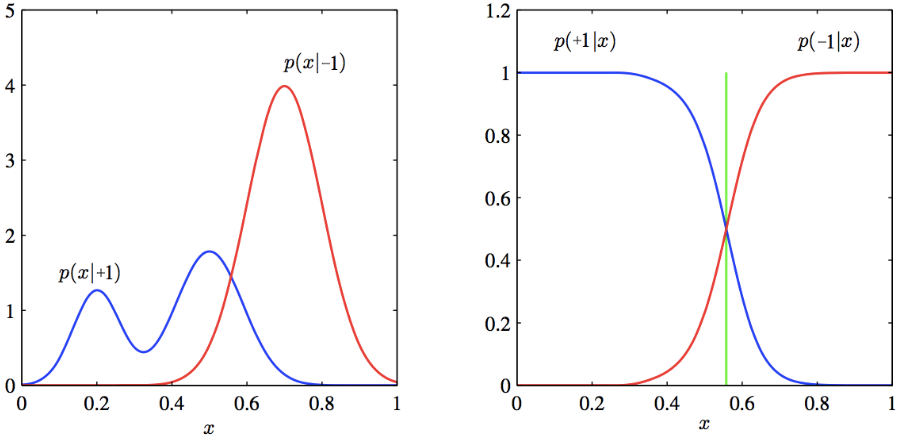
$$P(X, Y) = P(Y)P(X|Y) = P(X)P(Y|X).$$
$$P(X \in S \text{ and } Y = 1) = P(Y = 1)P(X \in S|Y = 1)$$





Detour: Joint and conditional distributions

$$\begin{aligned} P(X, Y) &= P(Y)P(X|Y) &= P(X)P(Y|X). \\ P(X \in S \text{ and } Y = 1) &= P(Y = 1)P(X \in S|Y = 1) &= P(X \in S)P(Y = 1|X \in S). \end{aligned}$$



Detour: Joint and conditional distributions

Bayes Theorem

Detour: Joint and conditional distributions

Bayes Theorem

$$P(Y = +1|X) =$$

Detour: Joint and conditional distributions

Bayes Theorem

$$P(Y = +1|X) = \frac{P(X|Y = +1)P(Y = +1)}{}.$$

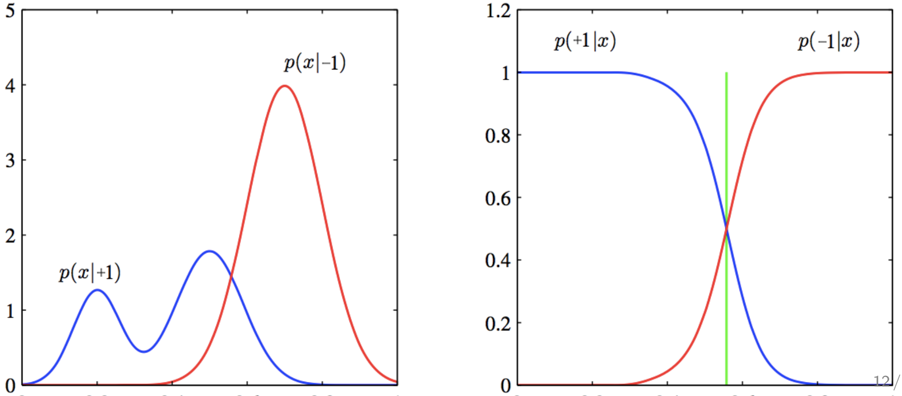
Detour: Joint and conditional distributions

Bayes Theorem

$$P(Y=+1|X) = \frac{P(X|Y=+1)P(Y=+1)}{P(X|Y=+1)P(Y=+1) + P(X|Y=-1)P(Y=-1)}.$$

Bayes Theorem

$$P(Y=+1|X) = \frac{P(X|Y=+1)P(Y=+1)}{P(X|Y=+1)P(Y=+1) + P(X|Y=-1)P(Y=-1)}.$$



Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (X, Y) pairs, $f(X)$ is a good prediction of Y.

1. Assume that (X_i, Y_i) and (X, Y) are chosen i.i.d. (independently and identically distributed), according to some probability distribution on $\mathcal{X} \times \mathcal{Y}$.

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (X, Y) pairs, $f(X)$ is a good prediction of Y.

1. Assume that (X_i, Y_i) and (X, Y) are chosen i.i.d. (independently and identically distributed), according to some probability distribution on $\mathcal{X} \times \mathcal{Y}$.
2. A *good prediction* means small expected loss

Probabilistic assumptions for prediction problems

The Prediction Problem

Given a *training set* of n pairs:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

choose a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for *subsequent* (X, Y) pairs, $f(X)$ is a good prediction of Y.

1. Assume that (X_i, Y_i) and (X, Y) are chosen i.i.d. (independently and identically distributed), according to some probability distribution on $\mathcal{X} \times \mathcal{Y}$.
2. A *good prediction* means small expected loss:
   The aim is to choose f with small *risk*,

$$R(f) = \mathbb{E}\ell(f(X), Y).$$

Risk

Risk

Example: Pattern classification

Risk

Example: Pattern classification

Risk is misclassification probability:

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) \qquad .$$

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}1[f(X) \neq Y] \qquad .$$

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

- Notation: Capital letters denote random variables.

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

- Notation: Capital letters denote random variables.
- The probability distribution models the relative frequency of different (X, Y) pairs.

Risk

Example: Pattern classification

Risk is misclassification probability:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

- Notation: Capital letters denote random variables.
- The probability distribution models the relative frequency of different (X, Y) pairs.
- It is crucial that the distribution of the training points (X_i, Y_i) is the same as that of the subsequent (X, Y) pair.

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk
  - **Bayes decision rule**
  - Excess risk
  - Risk, Bayes decision rule, excess risk in regression
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules

Minimizing Risk in Classification

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$R(f) = \mathbb{E}\ell(f(X), Y)$

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$$
\begin{aligned}
R(f) &= \mathbb{E}\ell(f(X), Y) \\
&= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X]
\end{aligned}
$$

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$$\begin{aligned} R(f) &= \mathbb{E}\ell(f(X), Y) \\ &= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] \\ &= \mathbb{E}\left[\ell(f(X), +1)P(Y = +1|X) + \right. \end{aligned}$$

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$$\begin{aligned} R(f) &= \mathbb{E}\ell(f(X), Y) \\ &= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] \\ &= \mathbb{E}\left[\ell(f(X), +1)P(Y = +1|X) + \ell(f(X), -1)P(Y = -1|X)\right] \end{aligned}$$

Minimizing Risk in Classification

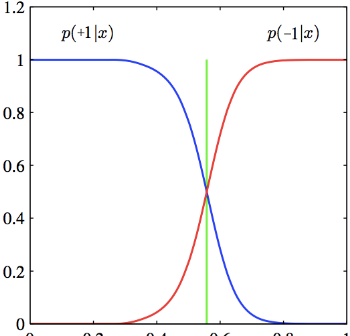
Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$$
\begin{aligned}
R(f) &= \mathbb{E}\ell(f(X), Y) \\
&= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] \\
&= \mathbb{E}\left[\ell(f(X), +1)P(Y = +1|X) + \ell(f(X), -1)P(Y = -1|X)\right] \\
&= \mathbb{E}\left[1[f(X) = -1]P(Y = +1|X) + 1[f(X) = +1]P(Y = -1|X)\right].
\end{aligned}
$$

Minimizing Risk in Classification

Two-class classification: $\mathcal{Y} = \{-1, +1\}$

$$
\begin{aligned}
R(f) &= \mathbb{E}\ell(f(X), Y) \\
&= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] \\
&= \mathbb{E}\left[\ell(f(X), +1)P(Y = +1|X) + \ell(f(X), -1)P(Y = -1|X)\right] \\
&= \mathbb{E}\left[1[f(X) = -1]P(Y = +1|X) + 1[f(X) = +1]P(Y = -1|X)\right].
\end{aligned}
$$



Minimizing Risk in Classification

Minimizing Risk in Classification

Bayes Decision Rule

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*.

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*. Denote the optimal risk (the *Bayes risk*), by

$$R^* = \inf_f R(f) \qquad .$$

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*. Denote the optimal risk (the *Bayes risk*), by

$$R^* = \inf_f R(f) = R(f^*).$$

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*. Denote the optimal risk (the *Bayes risk*), by

$$R^* = \inf_f R(f) = R(f^*).$$

If $P(Y = +1|x) = P(Y = -1|x) = 1/2$,

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*. Denote the optimal risk (the *Bayes risk*), by

$$R^* = \inf_f R(f) = R(f^*).$$

If $P(Y = +1|x) = P(Y = -1|x) = 1/2$, choice does not affect the risk. In that case, any choice for $f^*(x)$ is equally good. So there can be several Bayes decision rules.

Minimizing Risk in Classification

Bayes Decision Rule

Optimizing our choice for each X, we see that risk is minimized when $f = f^*$:

$$f^*(x) = \begin{cases} 1 & \text{if } P(Y = 1|x) > P(Y = -1|x), \\ -1 & \text{otherwise.} \end{cases}$$

This is called the *Bayes decision rule*. Denote the optimal risk (the *Bayes risk*), by

$$R^* = \inf_f R(f) = R(f^*).$$

If $P(Y = +1|x) = P(Y = -1|x) = 1/2$, choice does not affect the risk. In that case, any choice for $f^*(x)$ is equally good. So there can be several Bayes decision rules.

(How does f^* change if we have a different ℓ? c.f. the spam loss.)

Minimizing Risk in Classification

Minimizing Risk in Classification

Excess risk

Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

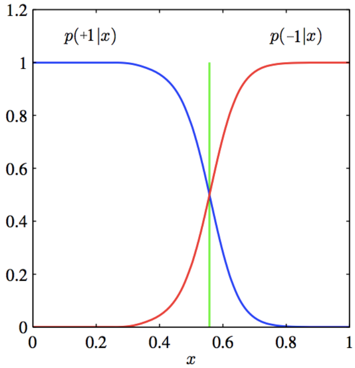
$$R(f) - R^* =$$

Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$R(f) - R^* =$$

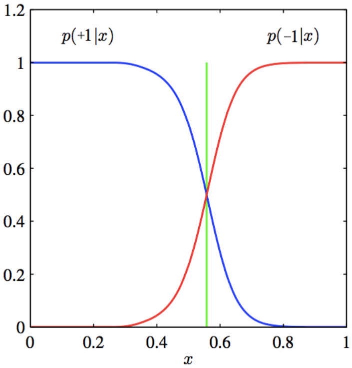


Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$R(f) - R^* = \mathbb{E}\left(1[f(X) \neq f^*(X)]\right]$$

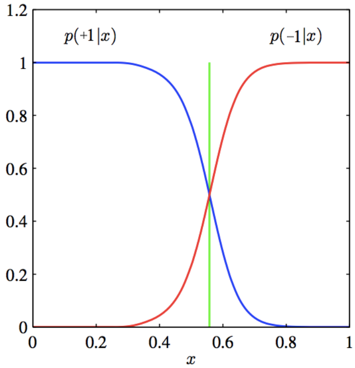


Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$R(f) - R^* = \mathbb{E}\left(1[f(X) \neq f^*(X)]|P(Y = +1|X) - P(Y = -1|X)|\right)$$

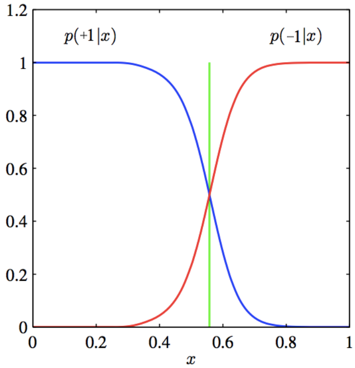


Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$\begin{aligned} R(f) - R^* &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|P(Y = +1|X) - P(Y = -1|X)|\right) \\ &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|2P(Y = +1|X) - 1|\right). \end{aligned}$$



Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$\begin{aligned} R(f) - R^* &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|P(Y = +1|X) - P(Y = -1|X)|\right) \\ &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|2P(Y = +1|X) - 1|\right). \end{aligned}$$

Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$R(f) - R^* = \mathbb{E}\left(1[f(X) \neq f^*(X)]|P(Y = +1|X) - P(Y = -1|X)|\right)$$
$$= \mathbb{E}\left(1[f(X) \neq f^*(X)]|2P(Y = +1|X) - 1|\right).$$

That is, the excess risk of a decision rule (above the Bayes risk) can be quantified in terms of a certain distance from f^*.

Minimizing Risk in Classification

Excess risk

For any $f : \mathcal{X} \to \{-1, +1\}$,

$$\begin{aligned} R(f) - R^* &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|P(Y = +1|X) - P(Y = -1|X)|\right) \\ &= \mathbb{E}\left(1[f(X) \neq f^*(X)]|2P(Y = +1|X) - 1|\right). \end{aligned}$$

That is, the excess risk of a decision rule (above the Bayes risk) can be quantified in terms of a certain distance from f^*.

(Not quite a distance: differences between functions at an x with $P(Y = +1|x) = 1/2$ have no influence on the risk.)

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk
  - Bayes decision rule
  - Excess risk
  - **Risk, Bayes decision rule, excess risk in regression**
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules

Risk

Example: Regression with squared loss

Risk

Example: Regression with squared loss

Risk is expected squared error:

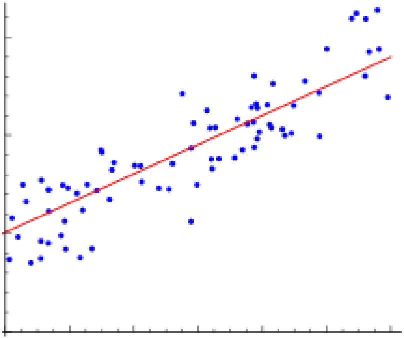
$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 .$$

Risk

Example: Regression with squared loss

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2\,.$$

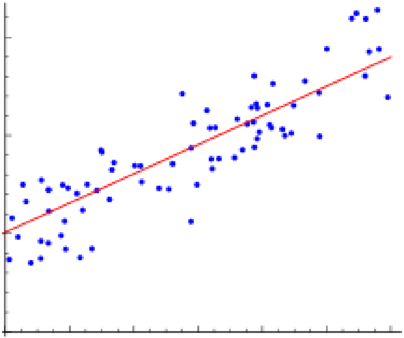


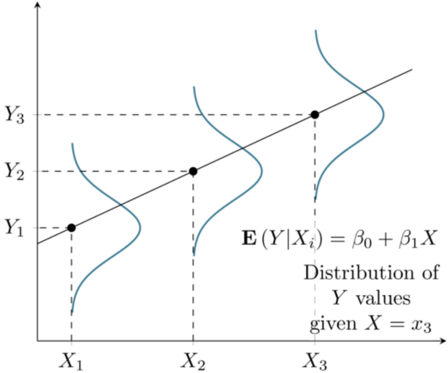
Risk

Example: Regression with squared loss

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2\,.$$





Risk in Regression

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2$$

Risk in Regression

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Risk in Regression

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Just as in the classification case, for each X, we minimize the conditional expectation of the loss,

$$\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Risk in Regression

Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Just as in the classification case, for each X, we minimize the conditional expectation of the loss,

$$\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Minimizing Risk in Regression

Bias-variance decomposition

Minimizing Risk in Regression

Bias-variance decomposition

$$R(f) = \mathbb{E}\,(f(X) - Y)^2$$

Minimizing Risk in Regression

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \mathbb{E}\left(f(X) - Y\right)^2 \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - Y\right)^2 | X\right]
\end{aligned}
$$

Minimizing Risk in Regression

Bias-variance decomposition

$$\begin{aligned} R(f) &= \mathbb{E}\left(f(X) - Y\right)^2 \\ &= \mathbb{E}\mathbb{E}\left[\left(f(X) - Y\right)^2 | X\right] \\ &= \mathbb{E}\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y\right)^2 | X\right] \end{aligned}$$

Minimizing Risk in Regression

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \mathbb{E}\left(f(X) - Y\right)^2 \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\Bigg[\left(f(X) - \mathbb{E}[Y|X]\right)^2 + \left(\mathbb{E}[Y|X] - Y\right)^2 \\
&\qquad + 2\left(f(X) - \mathbb{E}[Y|X]\right)\underbrace{\left(\mathbb{E}[Y|X] - Y\right)}\Bigg| X\Bigg]
\end{aligned}
$$

Minimizing Risk in Regression

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \mathbb{E}\left(f(X) - Y\right)^2 \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\Bigg[\left(f(X) - \mathbb{E}[Y|X]\right)^2 + \left(\mathbb{E}[Y|X] - Y\right)^2 \\
&\qquad + 2\left(f(X) - \mathbb{E}[Y|X]\right)\underbrace{\left(\mathbb{E}[Y|X] - Y\right)}\Bigg| X\Bigg] \\
&= \underbrace{\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X]\right)^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - Y\right)^2\right]}_{\text{variance}}
\end{aligned}
$$

Minimizing Risk in Regression

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \mathbb{E}\left(f(X) - Y\right)^2 \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y\right)^2 | X\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X]\right)^2 + \left(\mathbb{E}[Y|X] - Y\right)^2\right. \\
&\qquad\qquad \left. + 2\left(f(X) - \mathbb{E}[Y|X]\right)\underbrace{\left(\mathbb{E}[Y|X] - Y\right)}\bigg| X\right] \\
&= \underbrace{\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X]\right)^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - Y\right)^2\right]}_{\text{variance}}
\end{aligned}
$$

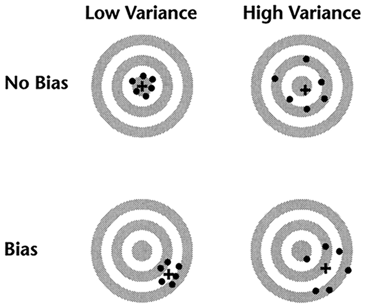
The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Minimizing Risk in Regression

$$R(f) = \mathbb{E}\underbrace{\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \mathbb{E}\underbrace{\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}.$$

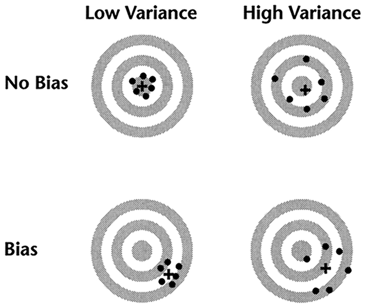
Minimizing Risk in Regression

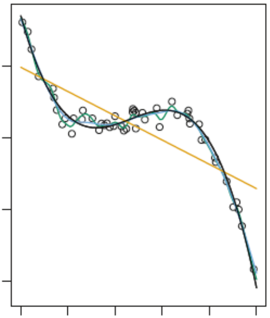
$$R(f) = \mathbb{E}\underbrace{\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \mathbb{E}\underbrace{\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}.$$



Minimizing Risk in Regression

$$R(f) = \mathbb{E}\underbrace{\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \mathbb{E}\underbrace{\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}.$$





Minimizing Risk in Regression

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

$$R(f) = \underbrace{\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}$$

Minimizing Risk in Regression

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

$$R(f) = \underbrace{\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}$$

$$= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - Y)^2\right]$$

Minimizing Risk in Regression

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

$$
\begin{aligned}
R(f) &= \underbrace{\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}} \\
&= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - Y)^2\right] \\
&= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + R\left(f^*\right).
\end{aligned}
$$

Minimizing Risk in Regression

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

$$R(f) = \underbrace{\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}$$

$$= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - Y)^2\right]$$

$$= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + R(f^*).$$

$$R(f) - R^* = \mathbb{E}\left[(f(X) - f^*(X))^2\right].$$

Detour: Bias and Variance of a Random Prediction Rule

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$.

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We'd like $\mathbb{E}R(\hat{f})$ to be small:

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We'd like $\mathbb{E}R(\hat{f})$ to be small:

$$\mathbb{E}R(\hat{f}) - R^* = \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right]$$

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We'd like $\mathbb{E}R(\hat{f})$ to be small:

$$\begin{aligned}\mathbb{E}R(\hat{f}) - R^* &= \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X) + \mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]\end{aligned}$$

Detour: Bias and Variance of a Random Prediction Rule

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We'd like $\mathbb{E}R(\hat{f})$ to be small:

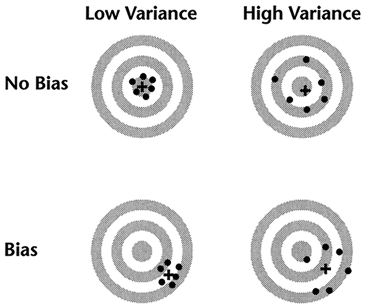
$$\begin{aligned}
\mathbb{E}R(\hat{f}) - R^* &= \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X) + \mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X)\right)^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]}_{\text{bias}^2}.
\end{aligned}$$

Detour: Bias and Variance of a Random Prediction Rule

$$\mathbb{E}R(\hat{f}) - R^* = \underbrace{\mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X)\right)^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]}_{\text{bias}^2}.$$

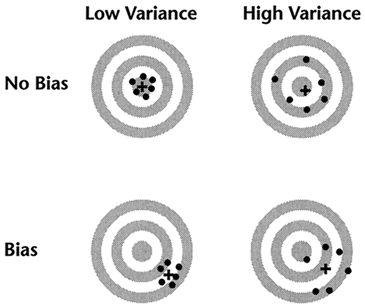
Detour: Bias and Variance of a Random Prediction Rule

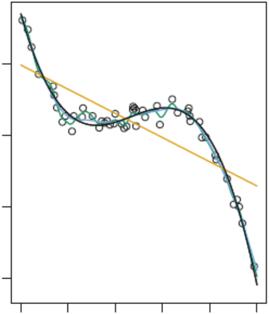
$$\mathbb{E}R(\hat{f}) - R^* = \underbrace{\mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X)\right)^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]}_{\text{bias}^2}.$$



Detour: Bias and Variance of a Random Prediction Rule

$$\mathbb{E}R(\hat{f}) - R^* = \underbrace{\mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X)\right)^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]}_{\text{bias}^2}.$$





Outline

- Decision theory
    - Loss functions
    - Probabilistic assumptions
    - Risk
    - Bayes decision rule
    - Excess risk
    - Risk, Bayes decision rule, excess risk in regression
- **Three approaches to estimating a classifier:**
  generative models, discriminative models, decision rules

Three approaches to choosing classifiers

Three approaches to choosing classifiers

1. Choose a classifier directly, based on optimization of some criterion.

Three approaches to choosing classifiers

1. Choose a classifier directly, based on optimization of some criterion. (c.f. perceptron algorithm, SVMs)

Three approaches to choosing classifiers

1. Choose a classifier directly, based on optimization of some criterion. (c.f. perceptron algorithm, SVMs)
2. Estimate a model for the joint probability distribution of (X, Y):

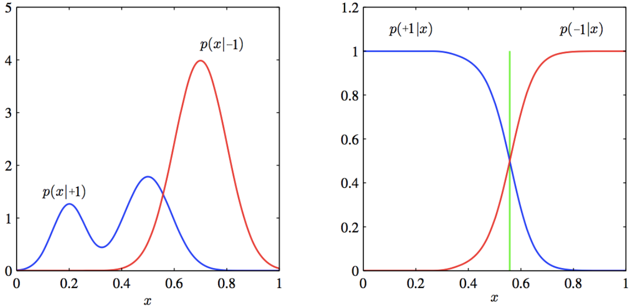
   and use it to construct a classifier.

Three approaches to choosing classifiers

1. Choose a classifier directly, based on optimization of some criterion. (c.f. perceptron algorithm, SVMs)
2. Estimate a model for the joint probability distribution of (X, Y):

   (a) Generative model $\quad P(X, Y) = P(Y)P(X|Y)$

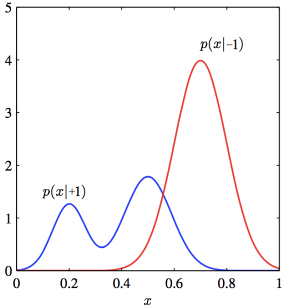
   and use it to construct a classifier.

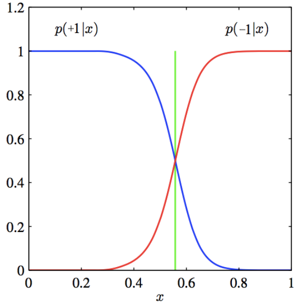


Three approaches to choosing classifiers

1. Choose a classifier directly, based on optimization of some criterion. (c.f. perceptron algorithm, SVMs)
2. Estimate a model for the joint probability distribution of (X, Y):
   - **(a)** Generative model $\quad P(X, Y) = P(Y)P(X|Y)$
   - **(b)** Discriminative model $\quad P(X, Y) = P(X)P(Y|X)$

   and use it to construct a classifier.





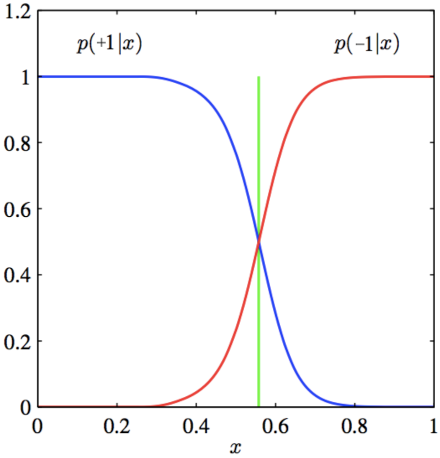
Three approaches to choosing classifiers

Discriminative models

Three approaches to choosing classifiers

Discriminative models

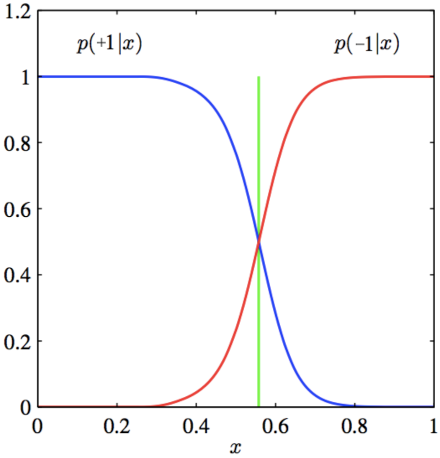
1. Estimate $P(Y|X)$.



Three approaches to choosing classifiers

Discriminative models

1. Estimate $P(Y|X)$.
2. Pretend that our estimate $\hat{P}(Y|X)$ is actually $P(Y|X)$ and substitute it in the expression for the Bayes rule:

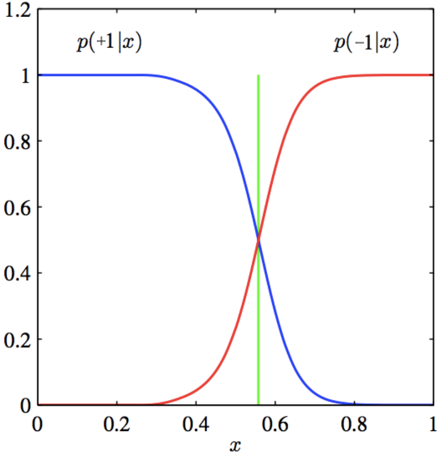


Three approaches to choosing classifiers

Discriminative models

1. Estimate $P(Y|X)$.
2. Pretend that our estimate $\hat{P}(Y|X)$ is actually $P(Y|X)$ and substitute it in the expression for the Bayes rule:

$$\hat{f}(x) = \begin{cases} 1 & \text{if } \hat{P}(Y=1|x) \\ & \quad > \hat{P}(Y=-1|x), \\ -1 & \text{otherwise.} \end{cases}$$



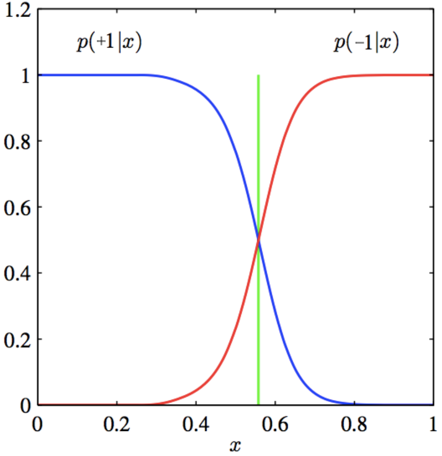
Three approaches to choosing classifiers

Discriminative models

1. Estimate $P(Y|X)$.
2. Pretend that our estimate $\hat{P}(Y|X)$ is actually $P(Y|X)$ and substitute it in the expression for the Bayes rule:

$$\hat{f}(x) = \begin{cases} 1 & \text{if } \hat{P}(Y=1|x) \\ & \quad > \hat{P}(Y=-1|x), \\ -1 & \text{otherwise.} \end{cases}$$

Called a *plug-in estimator*.



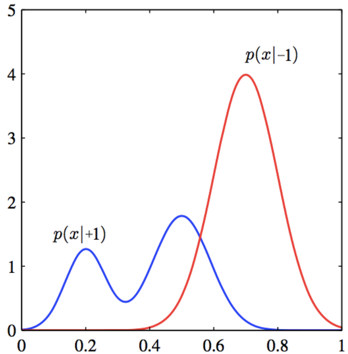
Three approaches to choosing classifiers

Generative models

Three approaches to choosing classifiers

Generative models

1. Estimate $P(Y)$ and $P(X|Y)$.

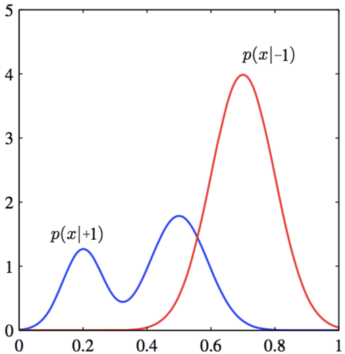


Three approaches to choosing classifiers

Generative models

1. Estimate $P(Y)$ and $P(X|Y)$.
2. Use Bayes theorem:

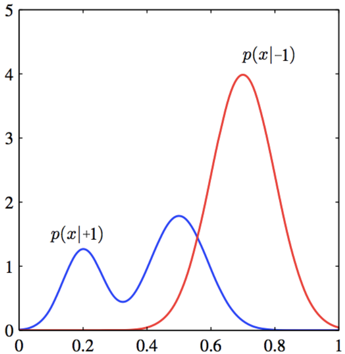
$$P(Y=+1|X) = \frac{P(X|Y=+1)P(Y=+1)}{P(X|Y=+1)P(Y=+1) + P(X|Y=-1)P(Y=-1)}.$$



Three approaches to choosing classifiers

Generative models

1. Estimate $P(Y)$ and $P(X|Y)$.
2. Use Bayes theorem:
$$P(Y=+1|X) = \frac{P(X|Y=+1)P(Y=+1)}{P(X|Y=+1)P(Y=+1) + P(X|Y=-1)P(Y=-1)}.$$
3. Define the plug-in estimator as for a discriminative model.



Three approaches to choosing classifiers

Three approaches to choosing classifiers

1. Estimate a generative model

Three approaches to choosing classifiers

1. Estimate a generative model

2. Estimate a discriminative model

Three approaches to choosing classifiers

1. Estimate a generative model

2. Estimate a discriminative model

3. Choose a classifier directly

Three approaches to choosing classifiers

1. Estimate a generative model:
   Estimate $P(X|Y)$.

2. Estimate a discriminative model

3. Choose a classifier directly

Three approaches to choosing classifiers

1. Estimate a generative model:
   Estimate $P(X|Y)$.
   But ultimately, all it uses is the conditional probability $P(Y|X)$.
2. Estimate a discriminative model
3. Choose a classifier directly

Three approaches to choosing classifiers

1. Estimate a generative model:
   Estimate $P(X|Y)$.
   But ultimately, all it uses is the conditional probability $P(Y|X)$.
2. Estimate a discriminative model:
   Directly estimate the conditional probability $P(Y|X)$.
3. Choose a classifier directly

Three approaches to choosing classifiers

1. Estimate a generative model:
   Estimate $P(X|Y)$.
   But ultimately, all it uses is the conditional probability $P(Y|X)$.
2. Estimate a discriminative model:
   Directly estimate the conditional probability $P(Y|X)$.
   But it typically aims to estimate it accurately across all values of X, when all that matters is whether $P(Y = +1|X) > P(Y = -1|X)$, so accuracy only matters where $P(Y = +1|X)$ is near $1/2$.
3. Choose a classifier directly

Three approaches to choosing classifiers

1. Estimate a generative model:
   Estimate $P(X|Y)$.
   But ultimately, all it uses is the conditional probability $P(Y|X)$.
2. Estimate a discriminative model:
   Directly estimate the conditional probability $P(Y|X)$.
   But it typically aims to estimate it accurately across all values of X, when all that matters is whether $P(Y = +1|X) > P(Y = -1|X)$, so accuracy only matters where $P(Y = +1|X)$ is near $1/2$.
3. Choose a classifier directly:
   By not solving a more difficult problem (e.g., density estimation), we might hope that this approach will do better.

Three approaches to choosing classifiers

This is not the whole story:

Three approaches to choosing classifiers

This is not the whole story:

- If we have a lot of information about the class-conditional distributions $P(X|Y)$, they might be much easier to estimate than conditionals or decision rules.

Three approaches to choosing classifiers

This is not the whole story:

- If we have a lot of information about the class-conditional distributions $P(X|Y)$, they might be much easier to estimate than conditionals or decision rules.
- If we have a lot of information about the conditional $P(Y|X)$, that might be informative about the decision boundary.

Three approaches to choosing classifiers

This is not the whole story:

- If we have a lot of information about the class-conditional distributions $P(X|Y)$, they might be much easier to estimate than conditionals or decision rules.
- If we have a lot of information about the conditional $P(Y|X)$, that might be informative about the decision boundary.
- Estimating a model can give extra information:

Three approaches to choosing classifiers

This is not the whole story:

- If we have a lot of information about the class-conditional distributions $P(X|Y)$, they might be much easier to estimate than conditionals or decision rules.
- If we have a lot of information about the conditional $P(Y|X)$, that might be informative about the decision boundary.
- Estimating a model can give extra information:
  e.g. an estimate of $P(Y=+1|X=x)$ conveys uncertainty.

Three approaches to choosing classifiers

This is not the whole story:

- If we have a lot of information about the class-conditional distributions $P(X|Y)$, they might be much easier to estimate than conditionals or decision rules.
- If we have a lot of information about the conditional $P(Y|X)$, that might be informative about the decision boundary.
- Estimating a model can give extra information:
  e.g. an estimate of $P(Y = +1|X = x)$ conveys uncertainty.
  e.g. an estimate of $P(X = x)$ can indicate if a point is an outlier.

Outline

- Decision theory
  - Loss functions
  - Probabilistic assumptions
  - Risk
  - Bayes decision rule
  - Excess risk
  - Risk, Bayes decision rule, excess risk in regression
- Three approaches to estimating a classifier:
  generative models, discriminative models, decision rules