Applying Statistical Decision Theory

Janos C. R. Füting

Department of Decision Sciences
HEC Montréal

May 27, 2024

The Elements of Statistical Decision Theory in Practice

Let's consider as an example the daily planning of a restaurant, that is:

- Have to plan for a guests tonight.
- Don't know true number of guests (θ).
- Do know how many came the last nights (X).

What is missing?

The Elements of Statistical Decision Theory in Practice

Let's consider as an example the daily planning of a restaurant, that is:

- Have to plan for a guests tonight.
- Don't know true number of guests (θ).
- Do know how many came the last nights (X).

What is missing?

The Elements of Statistical Decision Theory in Practice

Let's consider as an example the daily planning of a restaurant, that is:

- Have to plan for a guests tonight.
- Don't know true number of guests (θ).
- Do know how many came the last nights (X).

What is missing?

The Elements of Statistical Decision Theory in Practice

Let's consider as an example the daily planning of a restaurant, that is:

- Have to plan for a guests tonight.
- Don't know true number of guests (θ).
- Do know how many came the last nights (X).

What is missing?

Choosing our loss:

Let us say the cost of making an error is quadratic, i.e.

$$L(\delta(X), \theta) = \mathbb{E}\left[(\theta - \delta(X))^2\right]$$

Why?

- Larger errors are probably much worse in a lot of situations, while the direction matters less.
- Makes the math easy.
- Has an additional nice property that we can see ...

Choosing our loss:

Let us say the cost of making an error is quadratic, i.e.

$$L(\delta(X), \theta) = \mathbb{E}\left[(\theta - \delta(X))^2\right]$$

Why?

- Larger errors are probably much worse in a lot of situations, while the direction matters less.
- Makes the math easy.
- Has an additional nice property that we can see ...

Choosing our loss:

Let us say the cost of making an error is quadratic, i.e.

$$L(\delta(X), \theta) = \mathbb{E}\left[(\theta - \delta(X))^2\right]$$

Why?

- Larger errors are probably much worse in a lot of situations, while the direction matters less.
- Makes the math easy.
- Has an additional nice property that we can see ...

$$
\begin{aligned}
L(\delta(X),\theta) &= \mathbb{E}\left[(\theta-\delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{-\mathbb{E}[\delta(X)]+\mathbb{E}[\delta(x)]}_{0} -\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta-\mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)]-\delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2(\theta-\mathbb{E}[\delta(X)])\underbrace{(\mathbb{E}[\delta(X)]-\mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)]-\delta(X))^2] \\
&= Bias^2 \ + \ Variance
\end{aligned}
$$

Some Mathematical Magic

$$
\begin{aligned}
L(\delta(X),\theta) &= \mathbb{E}\left[(\theta-\delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{-\mathbb{E}[\delta(X)]+\mathbb{E}[\delta(x)]}_{0} -\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta-\mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)]-\delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2(\theta-\mathbb{E}[\delta(X)])\underbrace{(\mathbb{E}[\delta(X)]-\mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)]-\delta(X))^2] \\
&= Bias^2 \; + \; Variance
\end{aligned}
$$

Some Mathematical Magic

$$
\begin{aligned}
L(\delta(X),\theta) &= \mathbb{E}\left[(\theta-\delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{-\mathbb{E}[\delta(X)]+\mathbb{E}[\delta(x)]}_{0} -\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta-\mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)]-\delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2(\theta-\mathbb{E}[\delta(X)])\underbrace{(\mathbb{E}[\delta(X)]-\mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)]-\delta(X))^2] \\
&= Bias^2 \ + \ Variance
\end{aligned}
$$

$$
\begin{aligned}
L(\delta(X), \theta) &= \mathbb{E}\left[(\theta - \delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{- \mathbb{E}[\delta(X)] + \mathbb{E}[\delta(x)]}_{0} - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta - \mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)] - \delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2(\theta - \mathbb{E}[\delta(X)]) \underbrace{(\mathbb{E}[\delta(X)] - \mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)] - \delta(X))^2] \\
&= Bias^2 \; + \; Variance
\end{aligned}
$$

Some Mathematical Magic

$$
\begin{aligned}
L(\delta(X), \theta) &= \mathbb{E}\left[(\theta - \delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{-\mathbb{E}[\delta(X)] + \mathbb{E}[\delta(x)]}_{0} - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta - \mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)] - \delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2(\theta - \mathbb{E}[\delta(X)])\underbrace{(\mathbb{E}[\delta(X)] - \mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)] - \delta(X))^2] \\
&= Bias^2 \ + \ Variance
\end{aligned}
$$

Some Mathematical Magic

$$
\begin{aligned}
L(\delta(X),\theta) &= \mathbb{E}\left[(\theta-\delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{-\mathbb{E}[\delta(X)]+\mathbb{E}[\delta(x)]}_{0} -\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta-\mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)]-\delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + 2(\theta-\mathbb{E}[\delta(X)])\underbrace{(\mathbb{E}[\delta(X)]-\mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)]-\delta(X))^2] \\
&= (\theta-\mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)]-\delta(X))^2] \\
&= Bias^2 \; + \; Variance
\end{aligned}
$$

Some Mathematical Magic

$$
\begin{aligned}
L(\delta(X), \theta) &= \mathbb{E}\left[(\theta - \delta(X))^2\right] \\
&= \mathbb{E}[(\theta \underbrace{- \mathbb{E}[\delta(X)] + \mathbb{E}[\delta(x)]}_{0} - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2\mathbb{E}[(\theta - \mathbb{E}[\delta(x)])(\mathbb{E}[\delta(X)] - \delta(X))] + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + 2(\theta - \mathbb{E}[\delta(X)]) \underbrace{(\mathbb{E}[\delta(X)] - \mathbb{E}[\delta(X)])}_{0} + \mathbb{E}[(\mathbb{E}[\delta(x)] - \delta(X))^2] \\
&= (\theta - \mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\mathbb{E}[\delta(X)] - \delta(X))^2] \\
&= Bias^2 \; + \; Variance
\end{aligned}
$$

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

What does the Bias-Variance-Decomposition tell us?

Intuitively, the risk we are taking with a decision rule can be split into two kinds:

Bias: How wrong are we on average?

Variance: How far can our decisions be apart?

Sometimes we also call this fact the *Bias-Variance-tradeoff* because we may have two decision rules with the same overall risk, but a different balance between bias and variance.

Note that this only holds exactly for a quadratic loss, if we talk about this in other contexts we are refering to analogous tradeoffs.

The Bias Variance Tradeoff Visualised

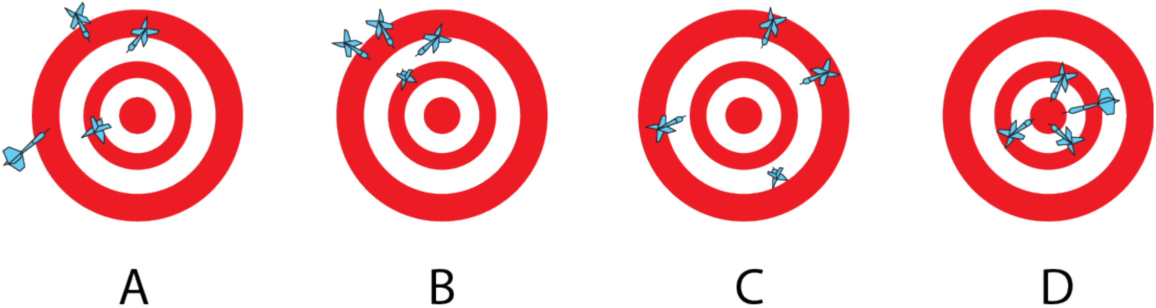

Figure: Image by Byron Inouye

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow$ 0, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow$ 0, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow$ 0, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

An Exercise

Two decision rules:

$\delta(X) = 0$ i.e. we always predict noone will show up ...

- Bias? $\rightarrow \theta$, we are wrong by whatever the true number is.
- Variance? $\rightarrow$ None, we always predict the same thing.

$\delta(X) = X_t$ i.e. we always predict the same number we had the night before

- Bias? $\rightarrow 0$, on average we are probably right.
- Variance? $\rightarrow$ Large, since we ignore all the other information we have.

Takeaways

What we have learned:

- Statistical Decision Theory tells us something about how to make and evaluate decisions when we face randomness.
- When our loss is quadratic we can decompose the decision risk into a bias component and a variance component.
- In some situations we can face a tradeoff between bias and variance when choosing between two decision rules.