MATH 829: Introduction to Data Mining and Analysis
Introduction to statistical decision theory

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

March 4, 2016

A framework for developing models. Suppose we want to predict a random variable Y using a random vector X.

Statistical decision theory

A framework for developing models. Suppose we want to predict a random variable Y using a random vector X.

- Let $\Pr(X, Y)$ denote the joint probability distribution of (X, Y).

Statistical decision theory

A framework for developing models. Suppose we want to predict a random variable Y using a random vector X.

- Let $\Pr(X, Y)$ denote the joint probability distribution of (X, Y).
- We want to predict Y using some function $g(X)$.

Statistical decision theory

A framework for developing models. Suppose we want to predict a random variable Y using a random vector X.

- Let $\Pr(X, Y)$ denote the joint probability distribution of (X, Y).
- We want to predict Y using some function $g(X)$.
- We have a *loss function* $L(Y, f(X))$ to measure how good we are doing, e.g., we used before

$$L(Y, f(X)) = (Y - g(X))^2.$$

when we worked with continuous random variables.

Statistical decision theory

A framework for developing models. Suppose we want to predict a random variable Y using a random vector X.

- Let $\Pr(X, Y)$ denote the joint probability distribution of (X, Y).
- We want to predict Y using some function $g(X)$.
- We have a *loss function* $L(Y, f(X))$ to measure how good we are doing, e.g., we used before

$$L(Y, f(X)) = (Y - g(X))^2.$$

  when we worked with continuous random variables.
- How do we choose g? "Optimal" choice?

Statistical decision theory (cont.)

Natural to minimize the *expected prediction error*:

$$\mathrm{EPE}(f) = E(L(Y, g(X))) = \int L(y, g(x)) \; \Pr(dx, dy).$$

Statistical decision theory (cont.)

Natural to minimize the *expected prediction error*:

$$\mathrm{EPE}(f) = E(L(Y, g(X))) = \int L(y, g(x)) \ \Pr(dx, dy).$$

For example, if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ have a *joint density* $f : \mathbb{R}^p \times \mathbb{R} \to [0, \infty)$, then we want to choose g to minimize

$$\int_{\mathbb{R}^p \times \mathbb{R}} (y - g(x))^2 f(x, y) \ dxdy.$$

Statistical decision theory (cont.)

Natural to minimize the *expected prediction error*:

$$\text{EPE}(f) = E(L(Y, g(X))) = \int L(y, g(x)) \ \Pr(dx, dy).$$

For example, if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ have a *joint density* $f : \mathbb{R}^p \times \mathbb{R} \to [0, \infty)$, then we want to choose g to minimize

$$\int_{\mathbb{R}^p \times \mathbb{R}} (y - g(x))^2 f(x, y) \ dxdy.$$

Recall the iterated expectations theorem:

- Let Z_1, Z_2 be random variables.
- Then $h(z_2) = E(Z_1 | Z_2 = z_2) =$ expected value of Z_1 w.r.t. the conditional distribution of Z_1 given $Z_2 = z_2$.
- We define $E(Z_1 | Z_2) = h(Z_2)$.

Now:

$$E(Z_1) = E\left(E(Z_1 | Z_2)\right).$$

Statistical decision theory (cont.)

Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\begin{aligned}\mathrm{EPE}(f) &= E\left[E[(Y - g(X))^2 | X]\right] \\ &= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x)\, dx.\end{aligned}$$

Statistical decision theory (cont.)

Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\begin{aligned}\mathrm{EPE}(f) &= E\left[E[(Y - g(X))^2 | X]\right] \\ &= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x)\; dx.\end{aligned}$$

Therefore, to minimize $\mathrm{EPE}(f)$, it suffices to choose

$$g(x) := \operatorname*{argmin}_{c \in \mathbb{R}} E[(Y - c)^2 | X = x].$$

Statistical decision theory (cont.)

Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\begin{aligned}\mathrm{EPE}(f) &= E\left[E[(Y - g(X))^2 | X]\right] \\ &= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x)\, dx.\end{aligned}$$

Therefore, to minimize $\mathrm{EPE}(f)$, it suffices to choose

$$g(x) := \operatorname*{argmin}_{c \in \mathbb{R}} E[(Y - c)^2 | X = x].$$

Expanding:

$$E[(Y - c)^2 | X = x] = E(Y^2 | X = x) - 2c \cdot E(Y | X = x) + c^2.$$

Statistical decision theory (cont.)

Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\begin{aligned}\text{EPE}(f) &= E\left[E[(Y - g(X))^2 | X]\right] \\ &= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x)\, dx.\end{aligned}$$

Therefore, to minimize $\text{EPE}(f)$, it suffices to choose

$$g(x) := \underset{c \in \mathbb{R}}{\text{argmin}}\, E[(Y - c)^2 | X = x].$$

Expanding:

$$E[(Y - c)^2 | X = x] = E(Y^2 | X = x) - 2c \cdot E(Y | X = x) + c^2.$$

The solution is

$$g(x) = E(Y | X = x).$$

Statistical decision theory (cont.)

Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\begin{aligned}\mathrm{EPE}(f) &= E\left[E[(Y - g(X))^2 | X]\right] \\ &= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x)\, dx.\end{aligned}$$

Therefore, to minimize $\mathrm{EPE}(f)$, it suffices to choose

$$g(x) := \operatorname*{argmin}_{c \in \mathbb{R}} E[(Y - c)^2 | X = x].$$

Expanding:

$$E[(Y - c)^2 | X = x] = E(Y^2 | X = x) - 2c \cdot E(Y | X = x) + c^2.$$

The solution is

$$g(x) = E(Y | X = x).$$

Best prediction: average given $X = x$.

Other loss functions

We saw that
$g(x) := \text{argmin}_{c \in \mathbb{R}} E[(Y - c)^2 | X = x] = E(Y | X = x)$.

Other loss functions

We saw that
$g(x) := \text{argmin}_{c \in \mathbb{R}} E[(Y-c)^2|X=x] = E(Y|X=x)$.

- Suppose instead we work with $L(Y, g(X)) = |Y - g(X)|$.

Other loss functions

We saw that
$g(x) := \operatorname{argmin}_{c\in\mathbb{R}} E[(Y-c)^2|X=x] = E(Y|X=x)$.

- Suppose instead we work with $L(Y, g(X)) = |Y - g(X)|$.
- Applying the same argument, we obtain

$$g(x) = \operatorname*{argmin}_{c\in\mathbb{R}} E\Big[|Y-c| \;\Big|\; X=x\Big].$$

Other loss functions

We saw that
$g(x) := \text{argmin}_{c\in\mathbb{R}} E[(Y-c)^2|X=x] = E(Y|X=x)$.

- Suppose instead we work with $L(Y, g(X)) = |Y - g(X)|$.
- Applying the same argument, we obtain

$$g(x) = \underset{c\in\mathbb{R}}{\text{argmin}}\, E[|Y-c| \;\big|\; X = x].$$

Problem: If X has density f_X, what is the min of $E(|X-c|)$ over c?

Other loss functions

We saw that
$g(x) := \text{argmin}_{c\in\mathbb{R}} E[(Y-c)^2|X=x] = E(Y|X=x)$.

- Suppose instead we work with $L(Y, g(X)) = |Y - g(X)|$.
- Applying the same argument, we obtain

$$g(x) = \underset{c\in\mathbb{R}}{\text{argmin}}\, E[|Y-c| \;\big|\; X=x].$$

Problem: If X has density f_X, what is the min of $E(|X-c|)$ over c?

$$\begin{aligned} E(|X-c|) &= \int |x-c|\ f_X(x)\ dx \\ &= \int_{-\infty}^{c} (c-x)\ f_X(x)dx + \int_{c}^{\infty} (x-c)\ f_X(x)dx. \end{aligned}$$

Other loss functions

We saw that
$g(x) := \text{argmin}_{c\in\mathbb{R}} E[(Y-c)^2|X=x] = E(Y|X=x)$.

- Suppose instead we work with $L(Y, g(X)) = |Y - g(X)|$.
- Applying the same argument, we obtain

$$g(x) = \underset{c\in\mathbb{R}}{\text{argmin}}\, E[|Y-c| \,\big|\, X=x].$$

Problem: If X has density f_X, what is the min of $E(|X-c|)$ over c?

$$\begin{aligned} E(|X-c|) &= \int |x-c|\ f_X(x)\ dx \\ &= \int_{-\infty}^{c} (c-x)\ f_X(x)dx + \int_c^{\infty} (x-c)\ f_X(x)dx. \end{aligned}$$

Now, differentiate

$$\frac{d}{dc}E(|X-c|) = \frac{d}{dc}\int_{-\infty}^{c} (c-x)\ f_X(x)dx + \frac{d}{dc}\int_c^{\infty} (x-c)\ f_X(x)dx$$

Other loss functions (cont.)

Recall:

$$\frac{d}{dx}\int_a^x h(t)\ dt = h(x).$$

Here, we have

$$\frac{d}{dc} c\int_{-\infty}^{c} f_X(x)dx - \int_{-\infty}^{c} x f_X(x)dx + \frac{d}{dc}\int_{c}^{\infty} x f_X(x)dx - c\int_{c}^{\infty} f_X(x)dx$$
$$= \int_{-\infty}^{c} f_X(x)dx - \int_{c}^{\infty} f_X(x)dx.$$

Check! (Use product rule and $\int_c^\infty = \int_{-\infty}^\infty - \int_{-\infty}^c$.)

Other loss functions (cont.)

Recall:

$$\frac{d}{dx}\int_a^x h(t)\ dt = h(x).$$

Here, we have

$$\frac{d}{dc}c\int_{-\infty}^{c} f_X(x)dx - \int_{-\infty}^{c} x f_X(x)dx + \frac{d}{dc}\int_c^{\infty} x f_X(x)dx - c\int_c^{\infty} f_X(x)dx$$
$$= \int_{-\infty}^{c} f_X(x)dx - \int_c^{\infty} f_X(x)dx.$$

Check! (Use product rule and $\int_c^\infty = \int_{-\infty}^{\infty} - \int_{-\infty}^{c}$.)

Conclusion: $\frac{d}{dc}E(|X-c|) = 0$ iff c is such that $F_X(c) = 1/2$. So the minimum of obtained when $c = \text{median}(X)$.

Other loss functions (cont.)

Recall:

$$\frac{d}{dx}\int_a^x h(t)\ dt = h(x).$$

Here, we have

$$\frac{d}{dc}c\int_{-\infty}^{c} f_X(x)dx - \int_{-\infty}^{c} xf_X(x)dx + \frac{d}{dc}\int_c^{\infty} xf_X(x)dx - c\int_c^{\infty} f_X(x)dx$$
$$= \int_{-\infty}^{c} f_X(x)dx - \int_c^{\infty} f_X(x)dx.$$

Check! (Use product rule and $\int_c^{\infty} = \int_{-\infty}^{\infty} - \int_{-\infty}^{c}$.)

Conclusion: $\frac{d}{dc}E(|X-c|) = 0$ iff c is such that $F_X(c) = 1/2$. So the minimum of obtained when $c = \mathrm{median}(X)$.

Going back to our problem:

$$g(x) = \operatorname*{argmin}_{c\in\mathbb{R}} E[|Y-c| \mid X = x] = \mathrm{median}(Y|X=x).$$

Back to nearest neighbors

We saw that $E(Y|X = x)$ minimize the expected loss with the loss is the squared error.

Back to nearest neighbors

We saw that $E(Y|X = x)$ minimize the expected loss with the loss is the squared error.

- In practice, we don't know the joint distribution of X and Y.

Back to nearest neighbors

We saw that $E(Y|X=x)$ minimize the expected loss with the loss is the squared error.

- In practice, we don't know the joint distribution of X and Y.
- The nearest neighbors can be seen as an attempt to approximate $E(Y|X=x)$ by
  1. Approximating the expected value by averaging sample data.
  2. Replacing "$|X=x$" by "$|X\approx x$" (since there is generally no or only a few samples where $X=x$).

Back to nearest neighbors

We saw that $E(Y|X = x)$ minimize the expected loss with the loss is the squared error.

- In practice, we don't know the joint distribution of X and Y.
- The nearest neighbors can be seen as an attempt to approximate $E(Y|X = x)$ by
  1. Approximating the expected value by averaging sample data.
  2. Replacing "$|X = x$" by "$|X \approx x$" (since there is generally no or only a few samples where $X = x$).

There is thus strong theoretical motivations for working with nearest neighbors.

We saw that $E(Y|X = x)$ minimize the expected loss with the loss is the squared error.

- In practice, we don't know the joint distribution of X and Y.
- The nearest neighbors can be seen as an attempt to approximate $E(Y|X = x)$ by
  1. Approximating the expected value by averaging sample data.
  2. Replacing "$|X = x$" by "$|X \approx x$" (since there is generally no or only a few samples where $X = x$).

There is thus strong theoretical motivations for working with nearest neighbors.

Note: If one is interested to control the absolute error, then one could compute the median of the neighbors instead of the mean.