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

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

March 11, 2020

The pmf/pdf of a random variable X:

- $f_X(x) = P(X = x)$ (discrete)
- $\int_A f_X(x)\, dx = P(X \in A)$ (continuous).

Review of probability theory

The pmf/pdf of a random variable X:

- $f_X(x) = P(X = x)$ (discrete)
- $\int_A f_X(x)\, dx = P(X \in A)$ (continuous).

Joint pmf/pdf of a random vector (X, Y):

- $f_{X,Y}(x, y) = P(X = x, Y = y)$ (discrete).
- $\iint_A f_{X,Y}(x, y)\, dxdy = P((X, Y) \in A)$ (continuous).

Review of probability theory

The pmf/pdf of a random variable X:

- $f_X(x) = P(X = x)$ (discrete)
- $\int_A f_X(x)\, dx = P(X \in A)$ (continuous).

Joint pmf/pdf of a random vector (X, Y):

- $f_{X,Y}(x, y) = P(X = x, Y = y)$ (discrete).
- $\iint_A f_{X,Y}(x, y)\, dxdy = P((X, Y) \in A)$ (continuous).

Expected value of a random variable:

- $E(X) = \sum_{i=1}^{N} x_i \cdot P(X = x_i)$ where $X \in \{x_1, \ldots, x_N\}$.
- $E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$.

Review of probability theory

The pmf/pdf of a random variable X:

- $f_X(x) = P(X = x)$ (discrete)
- $\int_A f_X(x)\, dx = P(X \in A)$ (continuous).

Joint pmf/pdf of a random vector (X, Y):

- $f_{X,Y}(x, y) = P(X = x, Y = y)$ (discrete).
- $\iint_A f_{X,Y}(x, y)\, dxdy = P((X, Y) \in A)$ (continuous).

Expected value of a random variable:

- $E(X) = \sum_{i=1}^{N} x_i \cdot P(X = x_i)$ where $X \in \{x_1, \ldots, x_N\}$.
- $E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$.

Expected value of a random vector $X = (X_1, \ldots, X_p)$ is
$E(X) = (E(X_1), E(X_2), \ldots, E(X_p))$.

Review of probability theory (cont.)

Marginal pmf/pdf:

- $f_X(x) = \sum_j f_{X,Y}(x, y_j)$ (discrete).
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ (continuous).

Review of probability theory (cont.)

Marginal pmf/pdf:

- $f_X(x) = \sum_j f_{X,Y}(x, y_j)$ (discrete).
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ (continuous).

Conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0.$$

Review of probability theory (cont.)

Marginal pmf/pdf:

- $f_X(x) = \sum_j f_{X,Y}(x, y_j)$ (discrete).
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ (continuous).

Conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0.$$

Conditional distributions:

- $f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ (discrete).
- $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ (continuous).

Marginal pmf/pdf:

- $f_X(x) = \sum_j f_{X,Y}(x, y_j)$ (discrete).
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ (continuous).

Conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0.$$

Conditional distributions:

- $f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ (discrete).
- $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ (continuous).

Conditional expectation:

- $E(X|Y = y_j) = \sum_i x_i \cdot P(X = x_i|Y = y_j) = \sum_i x_i \cdot f_{X|Y}(x_i|y_j)$.
- $E(X|Y = y) = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx$.

Review of probability theory (cont.)

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Review of probability theory (cont.)

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Example. If X is discrete, say $X \in \{x_1, \ldots, x_N\}$, then $Y = f(X)$ takes the value $f(x_i)$ with probability $P(X = x_i)$.

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Example. If X is discrete, say $X \in \{x_1, \ldots, x_N\}$, then $Y = f(X)$ takes the value $f(x_i)$ with probability $P(X = x_i)$.

Consider $f(y) = E(X|Y = y)$. This is a function. We define:

Review of probability theory (cont.)

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Example. If X is discrete, say $X \in \{x_1, \ldots, x_N\}$, then $Y = f(X)$ takes the value $f(x_i)$ with probability $P(X = x_i)$.

Consider $f(y) = E(X|Y = y)$. This is a function. We define:

$$E(X|Y) = f(Y).$$

This is a **random variable**.

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Example. If X is discrete, say $X \in \{x_1, \ldots, x_N\}$, then $Y = f(X)$ takes the value $f(x_i)$ with probability $P(X = x_i)$.

Consider $f(y) = E(X|Y = y)$. This is a function. We define:

$$E(X|Y) = f(Y).$$

This is a **random variable**.

Example. If Y is discrete, say $Y \in \{y_1, \ldots, y_M\}$, then $E(X|Y)$ takes the value $E(X|Y = y_i)$ with probability $P(Y = y_i)$.

Review of probability theory (cont.)

Recall: if X is a random variable and $f(\cdot)$ is some function, then $Y = f(X)$ is a new random variable.

Example. If X is discrete, say $X \in \{x_1, \ldots, x_N\}$, then $Y = f(X)$ takes the value $f(x_i)$ with probability $P(X = x_i)$.

Consider $f(y) = E(X|Y = y)$. This is a function. We define:

$$E(X|Y) = f(Y).$$

This is a **random variable**.

Example. If Y is discrete, say $Y \in \{y_1, \ldots, y_M\}$, then $E(X|Y)$ takes the value $E(X|Y = y_i)$ with probability $P(Y = y_i)$.

Theorem. (Iterated expectation theorem) We have

$$E(X) = E\left(E(X|Y)\right).$$

Review of probability theory (cont.)

Theorem. (Iterated expectation theorem) We have

$$E(X) = E\left(E(X|Y)\right).$$

Proof (discrete case). Suppose $X \in \{x_1, \ldots, x_N\}$ and $Y \in \{y_1, \ldots, y_M\}$. Then

$$\begin{aligned}
E(E(X|Y)) &= \sum_{j=1}^{M} E(X|Y = y_j) P(Y = y_j) \\
&= \sum_{j=1}^{M} \sum_{i=1}^{N} x_i \cdot P(X = x_i | Y = y_j) P(Y = y_j) \\
&= \sum_{j=1}^{M} \sum_{i=1}^{N} x_i \cdot \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} P(Y = y_j) \\
&= \sum_{j=1}^{M} \sum_{i=1}^{N} x_i \cdot P(X = x_i, Y = y_j) = \sum_{i=1}^{N} x_i \sum_{j=1}^{M} P(X = x_i, Y = y_j) \\
&= \sum_{i=1}^{N} x_i \cdot P(X = x_i) = E(X).
\end{aligned}$$

$\square$

Statistical decision theory

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 $f_{X,Y}(x,y)$ 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 $f_{X,Y}(x, y)$ denote the joint probability distribution of (X, Y).
- We want to predict Y using some function $g(X)$.

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

- Let $f_{X,Y}(x, y)$ 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.

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

- Let $f_{X,Y}(x, y)$ 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*:

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

Statistical decision theory (cont.)

Natural to minimize the *expected prediction error*:

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

For example, if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ have a *joint density* $f_{X,Y} : \mathbb{R}^p \times \mathbb{R} \to [0, \infty)$ and $L(x, y) = (x, y)^2$, then we want to choose g to minimize

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

Statistical decision theory (cont.)

Natural to minimize the *expected prediction error*:

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

For example, if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ have a *joint density* $f_{X,Y} : \mathbb{R}^p \times \mathbb{R} \to [0, \infty)$ and $L(x, y) = (x, y)^2$, then we want to choose g to minimize

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

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

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

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

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

Expanding:

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

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

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

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

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 $L(Y, g(X)) = |Y - g(X)|$.

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

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

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

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

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?

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

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

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

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

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

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

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 $c = \text{median}(X)$.

Other loss functions (cont.)

Recall:

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

Here, we have

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

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 $c = \mathrm{median}(X)$.

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.

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.

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