A framework for developing models. Suppose we want to predict a random variable $Y$ using a random vector $X$. 
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 $(X, Y)$. 
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 well 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?
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.
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?
Natural to minimize the *expected prediction error*.

\[
EPE(f) = E(L(Y, g(X))) = \int L(y, g(x)) \Pr(dx, dy).
\]
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

$$\int_{\mathbb{R}^p \times \mathbb{R}} (y - g(x))^2 f(x, y) \, dx \, dy.$$
Natural to minimize the *expected prediction error*.

\[
EPE(f) = E(L(Y, g(X))) = \int L(y, g(x)) \Pr(dx, dy).
\]

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

\[
\int_{\mathbb{R}^p \times \mathbb{R}} (y - g(x))^2 f(x, y) \, dx \, dy.
\]

**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:

\[
E(Z_1) = E(E(Z_1|Z_2)).
\]
Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\text{EPE}(f) = E \left[ E[(Y - g(X))^2 | X] \right]$$

$$= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x) \, dx.$$
Suppose \( L(Y, g(X)) = (Y - g(X))^2 \). Using the iterated expectations theorem:

\[
EPE(f) = E \left[ E[(Y - g(X))^2 | X] \right]
\]

\[
= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x) \, dx.
\]

Therefore, to minimize \( EPE(f) \), it suffices to choose

\[
g(x) := \arg\min_{c \in \mathbb{R}} E[(Y - c)^2 | X = x].
\]
Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$\text{EPE}(f) = E\left[ E[(Y - g(X))^2 | X] \right]$$

$$= \int E[(Y - g(X))^2 | X = x] \cdot f_X(x) \, dx.$$

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:

$$E[(Y - c)^2 | X = x] = E(Y^2 | X = x) - 2c \cdot E(Y | X = x) + c^2.$$
Suppose $L(Y, g(X)) = (Y - g(X))^2$. Using the iterated expectations theorem:

$$EPE(f) = E \left[ E[(Y - g(X))^2 | X] \right] = \int E[(Y - g(X))^2 | X = x] \cdot f_X(x) \, dx.$$  

Therefore, to minimize $EPE(f)$, it suffices to choose

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

Expanding:

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

The solution is

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

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

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

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

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 $X = x$. 
We saw that
\[ g(x) := \arg\min_{c \in \mathbb{R}} E[(Y - c)^2 | X = x] = E(Y | X = x). \]
We saw that
\[ g(x) := \arg\min_{c \in \mathbb{R}} E[(Y - c)^2 | X = x] = E(Y | X = x). \]
- Suppose instead we work with \( L(Y, g(X)) = |Y - g(X)| \).
Other loss functions

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

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

\[ g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| | X = x]. \]
We saw that
\[ g(x) := \arg\min_{c \in \mathbb{R}} E[(Y - c)^2 \mid X = x] = E(Y \mid X = x). \]

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

\[ g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| \mid X = x]. \]

**Problem:** If \( X \) has density \( f_X \), what is the min of \( E(|X - c|) \) over \( c \)?
Other loss functions

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

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

\[ g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| | X = x]. \]

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

\[
E(|X - c|) = \int |x - c| f_X(x) \, dx \\
= \int_{-\infty}^{c} (c - x) f_X(x) \, dx + \int_{c}^{\infty} (x - c) f_X(x) \, dx.
\]
Other loss functions

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

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

Applying the same argument, we obtain

\[ g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| | X = x]. \]

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

\[
E(|X - c|) = \int |x - c| \ f_X(x) \ dx
\]

\[
= \int_{-\infty}^{c} (c - x) \ f_X(x) \ dx + \int_{c}^{\infty} (x - c) \ f_X(x) \ dx.
\]

Now, differentiate

\[
\frac{d}{dc} E(|X - c|) = \frac{d}{dc} \int_{-\infty}^{c} (c - x) \ f_X(x) \ dx + \frac{d}{dc} \int_{c}^{\infty} (x - c) \ f_X(x) \ dx
\]
Recall:
\[
\frac{d}{dx} \int_{a}^{x} h(t) \, dt = h(x).
\]

Here, we have
\[
\frac{d}{dc} \int_{-\infty}^{c} f_X(x) \, dx - \int_{-\infty}^{c} x f_X(x) \, dx + \frac{d}{dc} \int_{c}^{\infty} x f_X(x) \, dx - c \int_{c}^{\infty} f_X(x) \, dx
\]
\[
= \int_{-\infty}^{c} f_X(x) \, dx - \int_{c}^{\infty} f_X(x) \, dx.
\]

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

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

Here, we have

$$\frac{d}{dc} \left( \int_{-\infty}^{c} f_X(x) \, dx - \int_{-\infty}^{c} x f_X(x) \, dx + \int_{c}^{\infty} x f_X(x) \, dx - c \int_{c}^{\infty} f_X(x) \, dx \right)$$

$$= \int_{-\infty}^{c} f_X(x) \, dx - \int_{c}^{\infty} f_X(x) \, dx.$$

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

\[
\frac{d}{dc} \left( \int_{-\infty}^c f_X(x) \, dx - \int_{-\infty}^c x f_X(x) \, dx + \int_0^c x f_X(x) \, dx \right) - c \int_0^c f_X(x) \, dx
\]

\[
= \int_{-\infty}^c f_X(x) \, dx - \int_c^\infty f_X(x) \, dx.
\]

Check! (Use product rule and \( \int_0^\infty = \int_{-\infty}^\infty - \int_{-\infty}^c \).)

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

Going back to our problem:

\[
g(x) = \arg\min_{c \in \mathbb{R}} \mathbb{E}[|Y - c| \mid X = x] = \text{median}(Y \mid X = x).
\]
We saw that $E(Y|X = x)$ minimize the expected loss with the loss is the squared error.
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 $Y$. 

Note: If one is interested to control the absolute error, then one could compute the median of the neighbors instead of the mean.
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 \( X = x \)).
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.