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:

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

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

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

Recall the iterated expectation 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:

\[
E(Z_1) = E(E(Z_1 | Z_2)).
\]

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

\begin{itemize}
  \item Suppose instead we work with \( L(Y, g(X)) = |Y - g(X)| \).
  \item Applying the same argument, we obtain
    \[ g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| | X = x] \].
\end{itemize}

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

### Back to nearest neighbors

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

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

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.

Other loss functions (cont.)

Recall:
\[
\frac{d}{dx} \int_{a}^{x} h(t) \, dt = h(x).
\]

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

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

Going back to our problem:
\[
g(x) = \arg\min_{c \in \mathbb{R}} E[|Y - c| | X = x] = \text{median}(Y | X = x).
\]