Example: image classification

Indoor

outdoor
Example: image classification (multiclass)

ImageNet figure borrowed from vision.stanford.edu
Multiclass classification

• Given training data \( \{(x_i, y_i): 1 \leq i \leq n\} \) i.i.d. from distribution \( D \)
  • \( x_i \in \mathbb{R}^d, y_i \in \{1,2,\ldots,K\} \)

• Find \( f(x): \mathbb{R}^d \rightarrow \{1,2,\ldots,K\} \) that outputs correct labels

• What kind of \( f \)?
Approach 1: reduce to regression

• Given training data \(\{(x_i, y_i): 1 \leq i \leq n\}\) i.i.d. from distribution \(D\)

• Find \(f_w(x) = w^T x\) that minimizes \(\hat{L}(f_w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2\)

• Bad idea even for binary classification

Reduce to linear regression; ignore the fact \(y \in \{1, 2, \ldots, K\}\)
Approach 1: reduce to regression

Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

Figure from *Pattern Recognition and Machine Learning*, Bishop
Approach 2: one-versus-the-rest

• Find $K - 1$ classifiers $f_1, f_2, \ldots, f_{K-1}$
  • $f_1$ classifies $1$ vs $\{2, 3, \ldots, K\}$
  • $f_2$ classifies $2$ vs $\{1, 3, \ldots, K\}$
  • \ldots
  • $f_{K-1}$ classifies $K - 1$ vs $\{1, 2, \ldots, K - 2\}$
  • Points not classified to classes $\{1, 2, \ldots, K - 1\}$ are put to class $K$

• Problem of ambiguous region: some points may be classified to more than one classes
Approach 2: one-versus-the-rest

Figure from *Pattern Recognition and Machine Learning*, Bishop
Approach 3: one-versus-one

- Find \((K - 1)K/2\) classifiers \(f_{(1,2)}, f_{(1,3)}, \ldots, f_{(K-1,K)}\)
  - \(f_{(1,2)}\) classifies 1 vs 2
  - \(f_{(1,3)}\) classifies 1 vs 3
  - ...
  - \(f_{(K-1,K)}\) classifies \(K - 1\) vs \(K\)

- Computationally expensive: think of \(K = 1000\)
- Problem of ambiguous region
Approach 3: one-versus-one

Figure from *Pattern Recognition and Machine Learning*, Bishop.
Approach 4: discriminant functions

• Find $K$ scoring functions $s_1, s_2, \ldots, s_K$
• Classify $x$ to class $y = \text{argmax}_i s_i(x)$

• Computationally cheap
• No ambiguous regions
Approach 4: discriminant functions

• Find $K$ discriminant functions $s_1, s_2, \ldots, s_K$

• Classify $x$ to class $y = \arg\max_i s_i(x)$

• Linear discriminant: $s_i(x) = (w^i)^T x$, with $w^i \in \mathbb{R}^d$
Approach 4: discriminant functions

• Linear discriminant: $s_i(x) = (w_i)^T x$, with $w_i \in R^d$
Approach 4: discriminant functions

• Linear discriminant: $s_i(x) = (w^i)^T x$, with $w^i \in \mathbb{R}^d$

• Lead to convex region for each class: by $y = \arg\max_i (w^i)^T x$
Approach 4: discriminant functions

• Find $K$ discriminant functions $s_1, s_2, \ldots, s_K$
• Classify $x$ to class $y = \text{argmax}_i s_i(x)$

• Conditional distributions: $s_i(x) = p(y = i|x)$
• Parametrize by $w^i: s_i(x) = p_{w^i}(y = i|x)$
Binary logistic regression: review

• Sigmoid

\[ \sigma(w^T x + b) = \frac{1}{1 + \exp(-(w^T x + b))} \]

• Interpret as conditional probability

\[ p_w(y = 1|x) = \sigma(w^T x + b) \]
\[ p_w(y = 0|x) = 1 - p_w(y = 1|x) = 1 - \sigma(w^T x + b) \]

• How to extend to multiclass?
Binary logistic regression: review

• Suppose we model the class-conditional densities $p(x|y = i)$ and class probabilities $p(y = i)$.

• Conditional probability by Bayesian rule:

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x|y = 1)p(y = 1) + p(x|y = 2)p(y = 2)} = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

where we define

$$a := \ln \frac{p(x|y = 1)p(y = 1)}{p(x|y = 2)p(y = 2)} = \ln \frac{p(y = 1|x)}{p(y = 2|x)}$$
Binary logistic regression: review

- Suppose we model the class-conditional densities \( p(x|y = i) \) and class probabilities \( p(y = i) \).

- \( p(y = 1|x) = \sigma(a) = \sigma(w^T x + b) \) is equivalent to setting log odds:
  \[
  a = \ln \frac{p(y = 1|x)}{p(y = 2|x)} = w^T x + b
  \]

- Why linear log odds?
Binary logistic regression: review

• Suppose the class-conditional densities $p(x|y = i)$ is normal

$$ p(x|y = i) = N(x|\mu_i, I) = \frac{1}{(2\pi)^{d/2}} \exp\left\{ -\frac{1}{2} \|x - \mu_i\|^2 \right\} $$

• log odd is

$$ a = \ln \frac{p(x|y = 1)p(y = 1)}{p(x|y = 2)p(y = 2)} = w^T x + b $$

where

$$ w = \mu_1 - \mu_2, \quad b = -\frac{1}{2} \mu_1^T \mu_1 + \frac{1}{2} \mu_2^T \mu_2 + \ln \frac{p(y = 1)}{p(y = 2)} $$
Multiclass logistic regression

• Suppose we model the class-conditional densities $p(x|y = i)$ and class probabilities $p(y = i)$

• Conditional probability by Bayesian rule:

$$p(y = i|x) = \frac{p(x|y = i)p(y = i)}{\sum_j p(x|y = j)p(y = j)} = \frac{\exp(a_i)}{\sum_j \exp(a_j)}$$

where we define

$$a_i := \ln [p(x|y = i)p(y = i)]$$
Multiclass logistic regression

• Suppose the class-conditional densities $p(x|y = i)$ is normal

\[
p(x|y = i) = N(x|\mu_i, I) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2} |x - \mu_i|^2 \right\}
\]

• Then

\[
a_i := \ln [p(x|y = i)p(y = i)] = -\frac{1}{2} x^T x + (w^i)^T x + b^i
\]

where

\[
w^i = \mu_i, \quad b^i = -\frac{1}{2} \mu_i^T \mu_i + \ln p(y = i)
\]
Multiclass logistic regression

• Suppose the class-conditional densities $p(x|y = i)$ is normal

$$p(x|y = i) = N(x|\mu_i, I) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}||x - \mu_i||^2\right\}$$

• Cancel out $-\frac{1}{2}x^T x$, we have

$$p(y = i|x) = \frac{\exp(a_i)}{\sum_j \exp(a_j)}, \quad a_i := (w^i)^T x + b^i$$

where

$$w^i = \mu_i, \quad b^i = -\frac{1}{2} \mu_i^T \mu_i + \ln p(y = i)$$
Multiclass logistic regression

• Suppose the class-conditional densities $p(x|y = i)$ is normal

$$p(x|y = i) = N(x|\mu_i, I) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2} ||x - \mu_i||^2 \right\}$$

• Then

$$p(y = i|x) = \frac{\exp( (w^i)^T x + b^i) }{\sum_j \exp( (w^j)^T x + b^j) }$$

which can be used to derive the negative log-likelihood (cross entropy)
Softmax

- A way to squash \( a = (a_1, a_2, ..., a_i, ...) \) into probability vector \( p \)

\[
\text{softmax}(a) = \left( \frac{\exp(a_1)}{\sum_j \exp(a_j)}, \frac{\exp(a_2)}{\sum_j \exp(a_j)}, ..., \frac{\exp(a_i)}{\sum_j \exp(a_j)}, ... \right)
\]

- Behave like max: when \( a_i \gg a_j (\forall j \neq i) \), \( p_i \approx 1, p_j \approx 0 \)
Cross entropy

• Let \( p_{data}(y|x) \) denote the empirical distribution of the data
• Negative log-likelihood
  \[
  -\frac{1}{n} \sum_{i=1}^{n} \log p(y = y_i|x_i) = -E_{p_{data}(y|x)} \log p(y|x)
  \]
  is the cross entropy between \( p_{data} \) and the model output \( p \)

• Information theory viewpoint: KL divergence
  \[
  D(p_{data}||p) = E_{p_{data}} [\log \frac{p_{data}}{p}] = E_{p_{data}} [\log p_{data}] - E_{p_{data}} [\log p]
  \]
  Entropy; constant       Cross entropy
Cross entropy

• Let $p_{\text{data}}(x, y)$ denote the empirical distribution of the data
• Negative log-likelihood

$$-\frac{1}{n} \sum_{i=1}^{n} \log p(x_i, y_i) = -\mathbb{E}_{p_{\text{data}}(x,y)} \log p(x, y)$$

is the cross entropy between $p_{\text{data}}$ and the model output $p$
Summary

Last hidden layer $h$

$$\begin{align*}
(w^j)^T h + b^j
\end{align*}$$

Softmax

Label $y_i$

Cross entropy

Linear

Convert to probability

Loss